STANFORD UNIVERSITY LIBRARY

THE CATHOLIC UNIVERSITY OF AMERICA

Measurements in the Fundamentals of Arithmetic

A Dissertation

Submitted to the School of Philosophy of the Catholic University of America, in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

BY

THOMAS GEORGE FORAN

STANFORD UNIVERSITY LIBRARIES
WITHDRAWN

The Catholic University of America
Washington, D. C.
1926

3781
C36F
UNCLASSED
Dissertation

THE CATHOLIC UNIVERSITY OF AMERICA

Measurements in the Fundamentals of Arithmetic

A Dissertation

Submitted to the School of Philosophy of the Catholic University of America, in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

BY

THOMAS GEORGE FORAN

The Catholic University of America
Washington, D. C.
1926

TABLE OF CONTENTS

		PAGE
INTRODUCTORY STATEMENT		5
CHAPTER I.	Principles Involved in Arithmetical Measurements	7
CHAPTER II.	A Survey of Current Computation Measurements	16
CHAPTER III.	Methods of Test and Scale Construction	39
CHAPTER IV.	The Diagnostic Computation Scale	46
CHAPTER V.	The Standardization of the Diagnostic Computation Scale	55
CHAPTER VI.	The Diagnostic Value of the Scale	63
BIBLIOGRAPHY		71

INTRODUCTORY STATEMENT

This study of arithmetic tests includes two parts. In the first the present status of such measurements is reviewed in some detail. Particular attention is directed towards the fundamental principles underlying tests and scales in computation. A large number of important tests have been critically examined in the light of these principles. The second part deals primarily with the derivation and standardization of a new scale for arithmetic fundamentals. The construction of this scale has been based on the principles of measurements as applied to current tests and on the results of experience with such tests. The aim has been to render the new scale as useful as possible in the situations for which tests and measurements have been devised. Many refinements that contribute but little to the usefulness of such tests or which actually detract from practical value have accordingly been omitted.

ACKNOWLEDGMENT

The writer wishes to acknowledge the valuable assistance and cooperation which many have rendered in the production of the Scales described in this study. He is particularly indebted to the Superintendents of Catholic Schools in whose dioceses the tests were given: Rev. J. F. Barbian, Milwaukee; Rev. T. V. Cassidy, Providence; Rev. F. J. Macelwane, Toledo; and Rev. U. J. Vehr, Cincinnati. To Sister M. Callixta, C.D.P., and the Sisters of St. Anne Convent, Melbourne, Ky., for their generous aid, the writer expresses his sincere appreciation. He gratefully acknowledges the assistance and encouragement which have been given him by Msgr. E. A. Pace, Dr. P. J. McCormick, Dr. T. V. Moore, and Dr. E. B. Jordan of the Catholic University. From these and many others he has received at all times invaluable and indispensable assistance.

CHAPTER I

Principles Involved in Arithmetical Measurements

As educational tests and measurements have developed, a number of requirements have been recognized with which such measurements must comply. The merits of particular tests and scales depend upon their conformity to these standards. These criteria necessarily influence the construction of educational measurements since their preparation, both in form and in content, conditions the results that are obtained from them. While many of these principles are generally accepted, the numerous educational tests available in any one subject, and the striking differences that exist between them, attest the differences of opinion regarding the desirable form and content of tests and scales. These variations in part but reflect the unsettled problems of other departments of education and particularly the identification of the objectives of instruction. Others, however, remain as illustrations and evidence of the immaturity of educational measurements and the absence of any generally accepted code of principles governing the construction of our tests and scales. Nor is it likely that any such general agreement will be effected within the near future. On the other hand, certain fundamental requirements are recognized by all and the variations in subject matter and in form are in the main due to differences in the more special requirements which depend on the particular subject of the curriculum. It is imperative that existing tests be carefully evaluated in order that those which fail to comply with reasonable demands be rejected, and that improvements in measurements as they appear be incorporated into new tests and scales.

There are a number of requirements or principles which are of a general type and which pertain to all educational measurements. Their identity is commonly agreed upon. In addition to them there are others of a more specialized nature which belong to tests and scales within but one subject. These are particular in their scope and application and do not extend beyond the tests of any one subject. The relative importance of these various requirements is difficult if not impossible to apportion. Indeed, any endeavor to decide this problem is in danger

of overlooking the necessity of all requirements being fulfilled. There is, however, probably a minimum essential degree existing in all below which a test ceases to be satisfactory. The amount of this necessary conformity is not known, and at best our standards for the evaluation of tests and measurements must be relative rather than absolute.

These general and specific requirements may be described with special reference to computation tests in general. The statistical methods employed in conjunction with these criteria need not be included within this statement of principles. In the succeeding chapter the standards may be applied to the more important measurements in computation. Those requirements that are most generally accepted constitute the basic principles of test evaluation and may accordingly be discussed first.

Of the general standards the first is the necessity for agreement between the content of a test or scale and the existing and accepted educational objectives in the subject with which it deals. Very frequently these aims and objectives are not clearly defined, and considerable latitude must be permitted in their interpretation. In some subjects we find almost complete recognition of certain objectives, while in others they are most indefinite and variable. Spelling affords an interesting illustration of the objectives within one subject. Here the social standard has forged to the front and the old aim of teaching the spelling of all words has been replaced by the determination of a comparatively limited number which form the core of the content of spelling. In arithmetic the social criterion is assuming primary importance. Both in problem solving and in computation many types of problems and of examples that formerly occupied prominent places are gradually but surely disappearing from texts and courses of study. While the exact aims of arithmetic teaching have not as yet been crystallized into a generally and completely accepted set of objectives, their present status has an important bearing upon the content of tests in computation as well as in problem solving. Some of these objectives may be stated negatively rather than positively. Very long multiplication examples, examples in cube root, and many other types are passing from the important position they so long usurped. If it is not desirable to teach such problems, still less is it desirable to include them within our tests and scales. There are,

nevertheless, a number of computation measurements in which such types and many others of a similar nature are included. In the identification of the content of the teaching of arithmetic, the work of Wilson[1] and that of Thorndike[2] have had extensive influence. Wilson, in particular, has shown that a very large number of the types of problems and of computation exercises, so frequently included in texts and in tests, have no justification for their inclusion that can be found in their social and business frequency. Any value that may be attached to them by virtue of their transfer power still awaits demonstration. Too frequently such examples displace others of greater importance and the merit of tests and scales which include them is correspondingly reduced. Conformity with educational objectives, therefore, may be stated as the primary general requirement of computation measurements.

A second principle involved in educational measurements is subsumed under the title of validity. Validity signifies the accurate representation by the scores of the abilities to be measured. Validity is reduced through the interference of other abilities than those primarily involved. Incidental abilities that may be invoked but which are not essentially related to the prime factor must always result in lack of validity. Computation tests should yield scores that are the direct product of arithmetical ability alone and uncomplicated by such extraneous factors as reading ability or rate of writing. Whenever large interfering factors are present, the resulting measurements cease to possess any real significance or to be accurate indices of those functions that the test purports to measure. Owing to the complexity of the mind and its operations and the interrelations of all abilities it is not possible to isolate completely specific abilities, but capacities other than those directly involved must be kept at a minimum and constant level. Tests and scales in computation vary in the extent to which various and miscellaneous capacities are controlled. As an illustration of this point, the influence which reading ability has upon performance

[1] Wilson, G. M.: "A Survey of the Social and Business Usage of Arithmetic." Teachers College Contributions to Education, No. 100, 1919, pp. vi+64.

[2] Thorndike, E. L.: "The Psychology of Arithmetic." New York: The Macmillan Co., 1922, pp. xvi+314.

in arithmetic may be cited. In tests that employ words to signify the operation to be performed for every example, the resulting scores will represent not only ability in computation but also and to an unknown extent the influence of the second variable, reading ability. Validity is a general standard of essential importance in all educational measurements. Whatever other merits an arithmetic test may possess, any departure from a high measure of validity renders it inefficient and at times dangerous to employ.

A third general standard is objectivity. Objectivity possesses two phases, the first pertaining to the administration of the test or scale and the second to the marking and scoring. Objectivity of administration is regulated by the clarity and completeness of the instructions supplied with the test. These instructions should be so framed that the possibility of misunderstanding or ambiguity is only remote. In its second aspect objectivity results from precise definitions of the methods of marking and of scoring. In its latter significance objectivity is comparatively easy to secure in arithmetic measurements. In all but a very small percentage of cases there can be no doubt regarding the correctness of a response.

Reliability constitutes a fourth general criterion for the evaluation of measurements. Reliability is employed in the restricted sense of agreement between the results of two applications of a test to the same pupils under parallel conditions. It is obvious that, if any test or scale renders different scores when given twice to the same pupils under identical conditions, such scores must be unreliable and untrustworthy. Many factors affect reliability, some of which are external to the tests and are less susceptible to influence while others are the product of the forces immediately connected with measurements themselves. In arithmetic tests surprisingly large variations are found in the reliability measures. Many of these are difficult to explain or account for. Low reliability vitiates a test and renders its significance in individual cases almost negligible.

In the construction and in the selection of tests and scales the facility with which they may be given and scored is an important practical consideration. This quality is the fifth of the general requirements. Ease of administration and of scoring is comparatively simple to secure and no valid reason has ever

been adduced for lengthy and cumbersome methods. Instructions should be specific and readily followed. Methods of scoring should be as simple as are consistent with reliable results. Intricate systems of scoring, so infrequently understood by those who employ tests, are handicaps and it is very difficult to demonstrate that their utility is in any way increased by their complexity. Too frequently statistical complications under the guise of scientific measurement conceal other limitations within the tests themselves. The present status of educational measurements fails to show either the advisability or necessity of any but simple methods of administration, marking and scoring.

A sixth principle involved in educational measurements deals with the norms or standards. Norms should be accurate and should be expressed in terms of both age and grade. The necessity for accurate norms requires no proof. In the case of computation tests that are primarily diagnostic rather than of the survey type, this requirement is lowered to some extent. Even in this class of tests, however, norms are necessary and therefore must be accurate. The application of the term "diagnostic" to a test should not be offered as an excuse for the absence or inadequacy of the norms. The deficiencies inherent in grade norms have been demonstrated in a number of occasions. They are suitable only for a particular time in the school year. Unless some system of derived scores be employed or corrections be available, results of tests given at other times are not comparable with any general standards. Of the various systems of derived scores, age scores appear to be the most readily understood and the most convenient for ordinary purposes. They lack some of merits of other derived scores but their more apparent meaning compensates for this inadequacy. Grades differ so greatly in the age and ability of the pupils that more refined methods are necessary than are afforded by the usual point scores. For these and other recognized reasons all educational measurements of the survey type should be provided with age scores and norms. Exceptions to this general rule exist in some diagnostic tests, in tests for the high school subjects, and in one, or possibly two subjects of the elementary school curriculum. In arithmetic tests there is a very real need of derived scores.

An eighth requirement of educational measurements demands duplicate forms for all such tests and scales. This requirement

is particularly cogent in arithmetic measurements. The essence of diagnostic tests and remedial teaching necessitates a measure of improvement. In instances where but one form is available re-tests lose a portion of their significance. It is therefore essential that a number of forms for all educational measurements be supplied. These duplicate forms must be of equivalent difficulty or the differences be known that allowance may be made for them. In the tool subjects of the elementary school a number of equivalent forms for all tests contribute materially to their usefulness. Many of the current tests in arithmetic possess but one form and their value is correspondingly limited.

A ninth and important practical consideration in the selection of tests and scales is their cost both in time and in money. When tests are to be repeated, classes are large, and several subjects are to be included, the financial outlay is a factor to be taken into consideration. The time involved in their administration and scoring is of similar importance. Ease of administration and of scoring has already been considered. Tests should be of sufficient length to insure their reliability, which is the paramount requirement, but reliability and facility of administration and of scoring are not incompatible. Measurements that require a long time to score cannot be used with any frequency on account of the labor that their employment exacts. At the other extreme are to be found tests where the time limits are so short as to place extremely high premiums on factors other than arithmetical ability alone. Their reliability is doubtful. But short time limits do not insure ease of scoring, for the larger number of tests and the greater complexity of the scoring in many such cases violate this principle. As in previous cases, the degree to which this requirement is fulfilled varies greatly with individual tests. Some computation measurements have such short time limits that an accurate sampling of the pupil's ability is precluded. A time limit of thirty seconds on such tests as these is wholly inadequate and permits gross errors of many types.

A tenth principle of measurements excludes the desirability of zero and perfect scores. Pupils who obtain such scores are unmeasured. The necessity of excluding zero scores amongst pupils possessing some degree of the ability involved arises from the failure of such scores to differentiate between the lower ex-

tremes of ability. At the upper limit of the scale, perfect scores will conceal true differences of ability that exist beneath apparently equal degrees of achievement. Arithmetic tests should extend to the lower degrees of ability in order that their diagnostic function may be utilized to the utmost and that faulty habits may be overcome before they become ingrained and difficult to eradicate. For this reason tests and scales in arithmetic that are not adapted to grades below the fourth are inadequate in their scope. The necessity for differential high scores is more important in other subjects than it is in arithmetic. The field of ability in computation is limited by the types of examples and the objectives of the subject. Mastery of the fundamentals will be revealed by perfect scores when the test is confined to basic abilities. Provided the test or scale includes all the important types of computation, perfect scores are not as undesirable as such scores would be in reading or other subjects where the limit of improvability is indefinite. In arithmetical computation, where the upper limit is defined by objectives, perfect scores do not violate the spirit of this principle if the scale from which they are derived conforms with such objectives. It is nearly always possible to construct a scale that will yield no zero scores but statistical appearance is secured at the sacrifice of other and more essential requirements of test and scale construction. Several computation scales have been criticized because of the frequency with which perfect scores are obtained. This criticism appears ill-directed and overlooks the essential nature of computation ability. A more serious limitation of certain scales is the presence of many zero scores among pupils who may be proved or presumed to possess some degree of the ability in question. Such a failure to record the lower degrees of computation ability lessens the value of such measurements very substantially. When perfect scores are obtained an examination of the content of the scale will reveal the general undesirability of such scores. In this respect spelling and arithmetic tests differ from those in reading where a still greater degree of ability is always possible and always measureable.

The foregoing requirements are wholly of a general type. Particular applications of these principles to arithmetical computation tests have been made in some instances. The remaining requirements are less extensive in their scope and are con-

cerned with computation tests alone. Concerning these there exists an absence of unanimity regarding the above-mentioned principles. In view of the many different types of tests and scales that exist, it will be inferred that the remaining requirements reflect mainly the opinions of a limited number of those engaged in the construction and evaluation of computation measurements.

The majority of arithmetic tests presume to be diagnostic to some extent. The purely survey tests in this subject constitute a small minority. But those that profess to be diagnostic differ radically in the degree to which this function is fulfilled. There appear to be wide differences in the interpretation of the term. In this study a diagnostic test or scale is understood as one that reveals the pupil's ability to solve various types of examples, the types with which he has difficulty, and the sources of his errors. Understood in this sense, the term "diagnostic" can correctly be applied to but a few of those that have adopted it. Some measurements merely show the fundamental operation in which the child is most deficient. Clearly these should not be designated as diagnostic for they include but one type of operation in the four fundamentals or a total of four in all. Others increase the number of operations to as many as twenty-one but the results that are obtained from such measurements are fundamentally measures of rate rather than an analysis of difficulties. They are essentially speed tests and fail to convey that information that may be legitimately demanded of diagnostic tests within the proper meaning of the term. It does not suffice to label a test as diagnostic, nor to give crude indications of only general types of disability, nor to reduce the analysis of causes to a measure of speed. A diagnostic test must reveal precise, not general, weaknesses and it must permit an analysis of their causes. It is maintained by many that the most important contribution of measurements to education is their capacity to reveal sources of disability. Tests and scales which fail in this respect are consequently secondary in value and importance to those that fulfill this requirement.

For the sake of convenience it is desirable that all the fundamental operations be included within one booklet, either through a spiral arrangement, in the case of speed tests, or by means of the mixed fundamental type of scale. Separate tests

in each of the four fundamentals, while serving an important purpose, are time consuming and unsatisfactory in certain respects. Those tests that are based upon the spiral arrangement should have a common unit of measurement for the convenient comparison of scores on the separate exercises. In current tests of this type, each exercise has a different unit and scores cannot be immediately compared. One correct answer in one exercise is seldom if ever equivalent to a correct answer in another exercise. Some system of equating the scores would enhance their significance and permit their comparison with ease. The desirability of common units extends also to separate scales for the four fundamental processes where a similar difficulty is encountered in the judgment of scores.

Differences of opinion exist regarding the desirable form of educational measurements. The development of tests and scales has rendered many of the older methods obsolete. It is evident that general agreement regarding the form of a computation scale is impossible at this stage. All forms of tests and scales possess some advantage while certain ones combine the merits of several types in one scale. A study of the present tests and scales in arithmetic has led to the opinion that speed tests, whether they claim to be diagnostic or not, are inferior in their general usefulness to those constructed on the mixed-fundamental plan.

A number of other requirements are so generally observed that reference to them is unnecessary. Allusions may be made to them in studying individual measurements in computation. This discussion of the requirements and principles involved in computation measurements will serve two purposes. These requirements may be conceived as criteria for the choice of tests and scales to employ. They may also be regarded as principles to be observed in the construction of new scales. In the former instance they have been utilized in the following chapter in order to evaluate current tests and scales, while in the latter they have been the determinants of the methods pursued in the construction of the Diagnostic Computation Scale.

CHAPTER II

A Survey of Current Computation Measurements

While at the present time a number of terms in educational measurements are used loosely and interchangeably, a more precise terminology is being evolved. The term "scale," hitherto employed to designate any form or type of educational measurement, is more correctly restricted to those measurements in which the items or examples increase in difficulty. In true scales the increments are equal and constant within the scale itself. In tests, properly so-called, all the items or examples are approximately equal in difficulty. A spiral test is a group of exercises, the items in any one exercise being of equivalent difficulty, but the exercises themselves becoming progressively more difficult. In order to distinguish between the several types of computation measurements, these terms will be employed only in their more exact meaning.

For purposes of discussion computation measurements may be classified into several types or categories. This classification, while not generally noted, serves to clarify the distinctions between types of measurements that are so important in evaluating the merits of the respective forms. Three main types may be distinguished. Of these the first type includes survey tests and scales, characterized by providing a single score to represent the individual's ability in general computation and by affording little or no opportunity for diagnosis of difficulties. This type of computation measurement has achieved but limited use and but few representatives of it are available. The second type includes those tests and scales that are primarily and essentially diagnostic, yielding separate scores for each of a series of exercises, but containing no provision for a total score for ability as a whole. At the present time this type predominates over the other two. Such tests as belong to this category are spiral in nature. In the third type of computation measurements survey and diagnostic functions are combined. Measurements of this class are all of the scale type. These may be further subdivided into scales for each of the fundamental operations, true scales, and mixed-fundamental scales. The survey function of such scales is provided through a single score that represents general

computation ability. An exception to this statement is to be found in the Woody scales that will be noted in due course. Their diagnostic function differs radically from that involved in the second type. Those tests that are termed diagnostic furnish a measurement of speed and a comparison of rates in a limited number of computation types. Diagnostic scales, on the other hand, emphasize methods and causes of errors with no measurement of speed being included except very indirectly. These types of measurements together with examples from each type may be summarized in the following manner:

- Computation measurements.
  - 1. Survey measurements.....
    - A. Tests, e. g., The Monroe General Survey.
    - B. Scales.
  - 2. Diagnostic tests.....
    - The Courtis Series B.
    - Monroe Diagnostic.
    - Cleveland Survey.
  - 3. Survey and diagnostic scales.
    - General scales: Woody A.
    - True Scales: Woody B.
    - Mixed Fundamental Scales: Woody-McCall, Stanford Computation, Pittsburgh Scales.

Some overlapping is unavoidable in the above schema. The Pittsburgh Scale is a true scale as well as a mixed fundamental scale. In fact, its inclusion in the latter class is doubtful but can be justified on certain bases. It might better be placed, perhaps, among the scales of the survey measurements were it not for the claim that it does yield a diagnostic measure. This problem will be considered in detail in connection with the scale itself.

Of the first type, survey measurements, the outstanding example is the Monroe General Survey Scale in Arithmetic.[3] This scale is more properly a spiral test. It has been derived largely from Monroe's Diagnostic Tests in Arithmetic and the Cleveland Survey Tests, both of which are essentially diagnostic tests. Monroe's General Survey Scale is available in two parts, Scale I for the third, fourth and fifth grades, and Scale II for the sixth, seventh and eighth grades. Two forms are published for each scale. In both cases separate scores are obtained for the various exercises comprising the scales. These scores are the

[3] Monroe, W. S.: "The Illinois Examination." University of Illinois, *Bureau of Educational Research, Bulletin No.* 6, 1921, p. 70.

weighted number of examples correct and the sum of the weighted scores yields a total point score which may be converted into age scores and B scores. Its survey function is fulfilled through this total point or derived score. While not professing to be diagnostic in nature, it may be used for this purpose to approximately the same extent as other speed tests that belong to the second type.

Scale I for the lower grades is formed of eight separate exercises, designated in the scale as Test 1, Test 2, etc. The content of the whole scale may be illustrated by examples drawn from Scale I, Form I.

Examples from Monroe's General Survey Arithmetic Scales

Test 1. Addition.

1	4	6	1	9
2	2	6	3	5

65 examples.
Time: 30 seconds.

Test 2. Subtraction.

9	12	7	1	11
9	3	3	0	6

50 examples.
Time: 30 seconds.

Test 3. Multiplication.

3	2	9	4	0
2	9	8	7	2

50 examples.
Time: 30 seconds.

Test 4. Division.

3)9 7)28 4)32 9)9

49 examples.
Time: 30 seconds.

Test 5. Addition.

5	6	2	1	6
2	3	8	4	8
2	4	8	2	0
0	8	5	5	7
4	8	1	4	6

32 examples.
Time: 1 minute.

Test 6. Subtraction.

37	94	60	27	39
5	8	3	6	7

32 examples.
Time: 1 minute.

Test 7. Multiplication.

6572	6750	5863	3754
6	9	2	5

20 examples.
Time: 1 minute.

Test 8. Division.

8)3840 4)7432 7)2534

20 examples.
Time: 2 minutes.

The equivalence of the examples within one test, the time limits which prevent perfect scores, the increasing difficulty of the separate tests, all stamp the Monroe General Survey Scale as a spiral test with the addition of a single score for the whole test. Its diagnostic properties are attained solely through the measurement of speed in the several operations. The scale for the upper grades is constructed on the same basis and differs only in the abilities demanded by the separate tests. Examples taken from Scale II, Form I, are given for purposes of illustration.

Examples from Monroe's General Survey Arithmetic Scales

SCALE II, FORM 1

Test 1. Addition.	7862 5013 1761 5872 3739	6809 7623 5299 6601 3496	8941 7910 9845 8522 1046
14 examples. Time: 4 minutes.			
Test 2. Multiplication.	4857 36	5718 92	6942 58
16 examples. Time: 4 minutes.			
Test 3. Division.	$41\overline{)574}$	$79\overline{)36893}$	$32\overline{)384}$
20 examples. Time: 4 minutes.			
Test 4. Subtraction.	739 367	1852 948	975 906
24 examples. Time: 1 minute.			
Test 5. Addition and subtraction of fractions. Addition: 8 examples. Subtraction: 7 examples. Time: 2 minutes.	$\frac{1}{6}+\frac{1}{3}=$	$\frac{3}{4}-\frac{2}{5}=$	$\frac{1}{6}+\frac{3}{5}=$
Test 6. Multiplication and division of fractions. Multiplication: 8 examples. Division: 7 examples. Time: 1½ minutes.	$\frac{2}{3}\times\frac{3}{4}=$		$\frac{4}{7}\div\frac{2}{3}=$
Test 7. Decimal fractions.	$.03\overline{)16.2}$	Ans.: 54	
	$.07\overline{)1.82}$	Ans.: 26	
	$.05\overline{).415}$	Ans.: 83	
24 examples. Time: 1 minute.			

An examination of these scales reveals their incompleteness. The whole field of computation is not covered nor are a sufficiently large number of types comprised within these tests to insure the total score representing general computation ability. This limitation is inherent in any test. The number of different abilities that comprise general computation skill is very large. Monroe has estimated that there are no less than 86 different abilities included.[4] No test can include exercises in all of these types without sacrificing many important requirements.

The standardization of these tests is undoubtedly adequate. In the derivation of age scores a large amount of exterpolation has been resorted to. The fact that this scale is not diagnostic in character seriously circumscribes its usefulness if measurements are to possess other than comparative functions. This criticism is directed against the entire class of tests and scales of this type. The Monroe General Survey is practically the only test that is essentially survey in character. If it can be demonstrated that survey and diagnostic functions can be combined in a single scale, such scales must necessarily be of greater value to teachers and to others than those possessing but a single function.

The second class of tests, the diagnostic, includes a larger number of measurements than does any other class. In general form tests belonging to this class are essentially similar to the Monroe Survey Scale but lack a total score for the entire test. Each separate test or exercise deals with a particular type of ability. The examples within an exercise are usually of equal difficulty. Time limits are used on all tests. More examples are included than it is possible for almost any pupil to answer in the time allowed. The score is the number of examples answered correctly. The score is, therefore, an index of speed.

The first of this class of tests is the Courtis Arithmetic Tests, Series B, available in three equivalent forms. This test contains four separate exercises, one in each of the four fundamentals. The structure and content of the test may be seen from the following examples:

[4] Monroe, Walter S.: "A Series of Diagnostic Tests in Arithmetic." *Elem. School J.*, xix, 1919, pp. 587–607.

Courtis Arithmetic Tests, Series B

Examples from Form 3

Test No. 1. Addition.

339	799	9
276	584	3
977	135	5
861	647	6
757	624	3
698	512	9
269	146	4
136	699	7
322	109	3

24 examples.
Time: 8 minutes.

Test No. 2. Subtraction

102142649	91885637
70428369	78379474

24 examples.
Time: 4 minutes.

Test No. 3. Multiplication.

9237	4568	6845
27	85	63

25 examples.
Time: 6 minutes.

Test No. 4. Division.

26)7306 74)66822 38)10640

24 examples.
Time: 8 minutes.

The adminstration of the Courtis tests entails no difficulties. Several systems of scoring are included, some of which are elaborate and cumbersome. The Courtis test was a pioneer in the field of educational measurements and has been extensively used. From the point of view of the principles discussed in the preceding chapter these tests show numerous and important defects. They cannot be used satisfactorily below the fourth grade and hence their scope is seriously limited. The unit of scoring in each of the four exercises is the example which renders the scoring extremely coarse. The addition problems of the first test each require twenty-seven operations. One error in these twenty-seven steps carries as much weight as ten or twenty such errors. With such coarse units there is great ambiguity in the interpretation of the scores. The difference between three examples correct and four examples correct is equal to the difference between eighty-one and 108 correct operations. The unit of scoring, therefore, becomes twenty-seven operations and the failure of the

test to distinguish between pupils who do not differ by as large a factor as this reduces the discriminative capacity of the test. A difference of one in the score on the addition exercise may represent a difference between eighty-one and 108 correct processes or the difference between 107 and 108 units. Several pupils who attain scores of five may in fact have performed from 135 to 161 correct operations. It is evident that such a system of scoring as this, which fails to give any credit for less than twenty-seven operations, may and will conceal wide differences. It is therefore unsatisfactory and a distinct source of errors in the interpretation of the results.

From the results secured by Woody[5] and by Guy[6] the reliability of the Courtis tests compares favorably with other tests of computation. Considered absolutely, however, the reliability is low. Coefficients of from 0.57 to 0.66 with probable errors of from 0.068 to 0.095 do not convey an impression of that degree of reliability that is desirable in educational measurements. In view of the low coefficients obtained on other tests and scales it might seem that there are factors inherent in computation ability itself that would cause such variability in the scores of children. Comparatively, however, the reliability of the Courtis tests is satisfactory. The size of the units undoubtedly has some effect on these measures.

While professing to be a diagnostic test, the Courtis reveals only the relative proficiencies in a few of the fundamental operations. They do not indicate specific difficulties, much less do they endeavor to analyze them. Kallom's statement that "The general knowledge that a pupil can or cannot add fractions is not sufficient," might be reasonably extended to cover the case of the Courtis tests.[7] Like all tests of its type, the Courtis exercises are measures of speed and the diagnosis, so-called, is not diagnosis within the precise meaning of the term.

The importance of the types of examples, particularly in addi-

---

[5] Woody, Clifford: "Results from Successive Repetitions of Certain Arithmetic Tests." Indiana University, *Bulletin of the School of Education,* i, No. 3, 1925, pp. 61–79.

[6] Guy, J. Freeman: "The Pittsburgh Arithmetic Scales." The University of Pittsburgh, pp. 47–49.

[7] Kallom, Arthur W.: "Analysis of and Testing in Common Fractions." *J. of Educ. Res.,* i, 1920, p. 187.

tion and subtraction, included in the Courtis tests may be seriously questioned. Wilson's study of the ordinary usages of arithmetic does not reveal the occurrence of any such types as Courtis has employed.[8] They are largely artificial from the standpoint of the social objectives. Moreover, before the child can reach the level of achievement demanded by the tests of the Courtis series, he must master the more elementary processes. These fundamental processes are the proper content of measurements in computation, and the failure to take them into account constitutes a fatal limitation to any test or scale. These several objections to the Courtis tests reveal many deficiencies in their form and content. While recognizing the influence that these tests have had and the value of the contributions they have made to measurements and to teaching, it may be concluded that the Courtis tests at the present time are inadequate and unsatisfactory for general usage.

Belonging to the same general type are the Monroe Diagnostic Tests in Arithmetic.[9] These tests are grouped in four parts and include in all twenty-one separate exercises arranged on the spiral plan. Tests 1–6 involve operations with integers, 7–11 more difficult operations with integers; tests 12–16 are tests in common fractions, while tests 17–21 include tests in the multiplication and division of decimal fractions. Each of the four parts is printed separately. The types of examples included may be seen from the following specimens taken from each of the twenty-one tests:

Examples From Monroe's Diagnostic Tests in Arithmetic

Test 1 Addition	Test 2 Subtraction	Test 3 Multiplication	Test 4 Division
4			
7	37	6572	
2	5	6	8)3840

[8] Wilson, G. M.: *loc. cit.*
Wilson, G. M., and Hoke, Kremer J.: "How to Measure." New York: The Macmillan Co., 1921, pp. vii+285.

[9] Monroe, W. S.: "A Series of Diagnostic Tests in Arithmetic." *Elem. School J.*, xix, 1919, pp. 585–607.

Test 5 Addition.	Test 6 Division.	Test 7 Addition	Test 8 Multiplication.
7862 5013 1761 5872 3739	82)3854	7 7 6 5 0 5 1 8	4857 36
Test 9 **Subtraction.**	**Test 10** **Multiplication.**	7 3 3	**Test 11** **Division.**
739 367	560 37	1 2	47)27589

Test 12 Addition.	Test 13 Subtraction.	Test 14 Multiplication.	Test 15 Addition.
$^1/_6+{}^1/_3=$	$^3/_4-{}^2/_5=$	$^2/_3\times{}^3/_4=$	$^1/_6+{}^3/_5=$

Test 16 Division	Test 17 Division	Test 18 Multiplication
$^2/_5\div{}^1/_3=$	.03)16.2 Ans.: 54	657.2 .7 46004

Test 19 Division	Test 20 Multiplication	Test 21 Division
.4)148 Ans.: 37	487.5 .62 302250	.47)2758.9

[1] In tests 17 to 21 inclusively, the answer is supplied, but the pupils are required to place the decimal point. Cf. Monroe, W. S., "The Ability to Place the Decimal Point in Division." *Elem. School J.*, xviii, 1917, pp. 287–293.

The objectivity of these tests is high. The time limits are short in a large number of tests; in seven of them only thirty seconds is allowed. This very short limit is a source of error in the scores, for a few seconds' inaccuracy in timing or the delay of pupils in commencing work promptly influences performance very substantially in such cases. An allowance of thirty seconds is insufficient on any educational test. These tests are, of necessity, speed tests. Accuracy is gauged by the proportion of correct examples to the number attempted. No norms are provided, however, for either accuracy or attempts. The system of scoring is similar to that employed by Courtis. Separate scores are obtained on each of the tests with no provision being made

for a total from all tests. Grade norms are available for grades four to eight, inclusively. In this respect the range of the tests is limited similarly to that of the Courtis. There is adequate opportunity and imperative need for measurements in computation below the fourth grade and tests that are not adapted to this requirement are thereby reduced in general value. Such tests must yield place to others that are suitable for the measurement of ability and the diagnosis of disability in the lower grades. The scoring is comparatively coarse and the differences between the grade norms are frequently negligible, thus hindering the comparative uses of the tests and their measurement of rate of work.

As is implied by their name, the Monroe Diagnostic Tests are intended for this function alone. The type of diagnosis is radically different from that possible through the use of certain types of scales. The scores are speed scores. Their only significance is the rate at which pupils can solve certain types of examples within a specified time. Practically no account is taken of the pupils' methods of work, of correct or incorrect habits, or of causes and sources of errors. The latter are obscured. Speed is necessarily dependent on the prior abilities to add, subtract, multiply and divide. The development of speed must be secondary. In measurements the same sequence should apply. When accuracy and power have been attained, the Monroe Diagnostic Tests will prove useful in the measurement of rate. The name that these tests bear should not confuse the fundamental issues involved nor lead to mistaken conceptions regarding the nature of diagnosis and its place and importance in arithmetical tests and scales. Certain dangers are associated with the accentuation of results obtained from such tests as these. Emphasizing speed scores detracts from the attention due to accuracy. The comparatively large improvement that may be secured through drill on speed is interpreted as evidence of greater computation ability and the distinction between the two variables is lost to view. Such large improvements imply no increase in the ability to solve more difficult types. It is therefore essential that the Monroe Diagnostic Tests and all such types of tests as these be subordinated to those that emphasize power and permit the diagnosis of incorrect methods and faulty habits rather than merely of deficiency in speed.

Only one form of these tests is available. Re-testing to determine improvements is therefore liable to gross practice influences. The usefulness of such tests as these should not, however, be overlooked despite the general deficiencies that have been discussed in some detail. With power tests these tests form a useful combination and their function and value are very real.

Fundamentally similar to the Monroe Diagnostic Tests are the Cleveland Survey Arithmetic Tests. The term *survey* in this connection is not to be confused with the meaning that has been attached to it in the classification of tests that was formed. The Cleveland tests are all included within one folder. They have been arranged on a spiral plan of increasing difficulty of the exercises. Fifteen separate exercises comprise the whole test. The nature of the performances required may be seen from examples taken from these tests.

Examples From the Cleveland Survey Arithmetic Tests

Set A. Addition.	Set B. Subtraction.	Set C. Multiplication.	Set D. Division.
1	7	3	
2	3	2	3)9

Set E. Addition.	Set F. Subtraction.	Set G. Multiplication.	Set H. Fractions.
5			
2			
2			
0	616	2345	
4	456	2	$^3/_5+^1/_5=$ $^6/_9-^4/_9=$

Set I. Division.	Set J. Addition.	Set K. Division.	Set L. Multiplication.
	7		8246
4)55424	5	21)273	29
	4		
	2		
	6		
Set M. Addition.	0	Set N. Division.	
	5		Set O. Fractions.
	1		
7493	8		
9016	3	67)32763	
6487	2		
7591	9		$^{15}/_{18}+^1/_6=$ $^9/_{14}-1/4=$
6166	2		

The fifteen sets of the Cleveland tests are not as extensive in their application as the twenty-one separate tests of the Monroe

Diagnostic series. They are speed tests. The time limits are short and in five sets are but thirty seconds each. As has been previously stated this allowance is entirely too short for valid results. It would have been preferable to lengthen the sets and to have doubled the time limits. In spite of these short limits the whole test consumes too much time to be completed at one period. The printing is by no means clear while many of the examples are so crowded together that pupils lack space to record their answers. The Cleveland tests are not intended for use below the fourth grade and therefore suffer from the same limitation that characterizes the other tests of this type. But one form of the tests has been devised. The norms are stated in terms of the median scores by grades of the schools of several large cities. These medians differ considerably for the cities involved and the interpretation of scores is thereby rendered most difficult and unsatisfactory. Norms are most essential for such tests as these and the absence of any general norms constitutes a deficiency that seriously militates against their usage. Their diagnostic value is equal to or less than that of other tests of the same general type. While belonging to the same class of measurements as the Monroe Diagnostic tests, they are by no means as satisfactory in their form, content, standardization, or convenience.

The Courtis Series B tests, the Monroe Diagnostic Tests and the Cleveland Survey tests are the principal representatives of those computation measurements included within the second class. Certain limitations that they possess have been defined. They are alike defective in failing to render that information that will permit teachers to localize sources of errors in computation. Their scores are measures of rate of computation rather than of power. All of them are restricted to grades from the fourth to the eighth. The absence of any general score makes them useless for survey purposes just as the diagnostic power of the purely survey tests is negligible. If it can be demonstrated that tests and scales can combine both these functions, those that possess but one must necessarily be of secondary usefulness for general purposes. The measurement of speed is essential but it must not be confused with nor supersede the measurement of power.

The third type of computation measurements includes those

that combine the survey and diagnostic functions within a single scale. An essential difference existing between such scales and diagnostic tests is the failure of the former to take into account rate of work. In scales, the measurement of accuracy and power replaces the measurement of speed in the second type of tests. If the analysis furnished by tests of the second type is a speed diagnosis, that of the combination scales is of power and of sources of disability. The latter is more elementary and more necessary. The scales belonging to the third type differ considerably among themselves. Some are separate scales for each of the four fundamental processes; others are true scales for each process; while others are of the mixed-fundamental type.

Woody[10] has constructed two series of scales, designated as Series A and Series B. The latter constitute true scales. Those of Series A are four in number, one for each of the fundamentals, in which the examples are arranged in ascending order of difficulty but are not equally spaced. A few examples may be selected from each of the Series A scales to illustrate them.

Woody Addition Scale, Series A

(1)	(7)	(13)	(19)	(25)
		23	$.75	
2		25	1.25	
3	3+1=	16	.49	$\frac{3}{8}+\frac{5}{8}+\frac{7}{8}+\frac{1}{8}=$

(31)	(37)
113.46	
49.6097	
19.9	
9.87	$16\frac{1}{3}$
.0086	$12\frac{1}{8}$
18.253	$21\frac{1}{2}$
6.04	$23\frac{3}{4}$

Examples From Woody Subtraction Scale, Series A

(1)	(6)	(11)	(16)	(21)
8	11	76	270	10.00
5	7	60	190	3.49

(26)	(31)	(35)
4 yds. 1 ft. 6 in.		
2 yds. 2 ft. 3 in.	7.3−3.00081=	$3\frac{7}{8}-1\frac{5}{8}=$

[10] Woody, Clifford: "Measurements of Some Achievements in Arithmetic." Teachers College, Columbia University, *Contributions to Education*, 1920, No. 80, p. 78.

Examples From Woody Multiplication Scale, Series A

(1)	(8)	(15)	(22)	(29)
	50	235		
$3\times7=$	3	23	$8\times5\frac{3}{4}=$	$\frac{1}{8}\times2=$

(36)	(38)
3 ft. 5 in.	.0963⅛
5	.084

Examples From Woody Division Scale, Series A

(1)	(6)	(11)	(16)	(21)
3)6	3)39	2)13	68)2108	25)9750

(26)	(31)	(36)
12)2.76	$\frac{5}{4}\div\frac{3}{5}=$	9)69 lbs. 9 oz.

As may be seen from these examples, the range of difficulty is considerable. Practically all important types are represented. A separate score is obtained for each of the scales but no method has been offered for the determination of a total score for all four scales. The time limit on each scale is twenty minutes. They may be used in all grades from the third to the eighth, inclusively. But one form for each of the scales has been published. Norms based on large numbers of cases are given in terms of grades. The correction of the scales is easily and quickly accomplished. Two methods of scoring accompany them. Of these one is comparatively difficult and is based upon the difficulty of the examples that the pupil solves correctly. On account of this comparatively cumbersome method, recourse is usually had to the simpler method of taking the number of examples correctly answered as the pupil's score. There is no evidence to show that the former method possesses any advantages that will compensate for its greater difficulty. In fact, the long method of scoring has been one of the objections that have been advanced in regard to these scales. The absence of duplicate forms is a limitation when retesting is desirable. If all four scales are employed, the time involved in their administration and scoring is apt to be greater than that usually devoted to tests and scales in computation. The greater accuracy of the results and the increased diagnostic value of these scales, however, compensate for this.

The diagnostic features of the Woody Scales have been a subject of some debate. Though not originally intended for this purpose, experience has shown that they contribute to this requirement in a very large degree. The large number of examples, covering practically the entire range of computation ability, permits the localization of those types of examples that are most difficult for a particular child or for a class. Not only may the relative difficulty of types be found but the methods which the children employ may be ascertained and thereby made the basis of remedial teaching. This feature is one that is entirely neglected by all tests that have been hitherto examined. It represents a use of educational measurements that may justifiably be regarded as their principal one. It is not merely a measure of speed such as is secured on the Cleveland tests, but a definition of difficulties and their causes. Since the ability to add, subtract, multiply or divide must necessarily exist before the child can add, subtract, multiply or divide rapidly, the functions involved in these scales are more important to develop than those in rate tests. The results obtained by Theisen and Flemming are conclusive answers to the argument of Monroe in favor of the Cleveland tests.[11] The method of employing such scales as these for diagnostic purposes will be considered in detail in a subsequent chapter.

In the Series B scales the examples have been so selected from those included in the A scales that they increase in difficulty by constant amounts. The B scales, therefore, are true scales. They are necessarily much shorter, having but approximately one-half as many examples as the longer tests, and require but half the time for their administration. The fewer number of examples, however, restricts their usefulness and counteracts the advantage of the shorter time limits. As in the case of the A scales, there is but one form. Their survey function is perhaps more important than their diagnostic power. McCall's results indicate that their reliability is so low as to render them applicable only to class rather than individual purposes.[12] The Series B scales are by no means as satisfactory as the A scales and the greater reli-

[11] Theisen, W. W., and Flemming, Cecil W.: "The Diagnostic Value of the Woody Arithmetic Scales: A Reply." *J. of Educ. Psy.*, ix, 1918, pp. 475–488, 567–580.

[12] McCall, William A.: "Measuring the Horace Mann Elementary School." *Teachers College Record*, xix, 1918, pp. 472–484.

ability and general value of the results obtained from the latter justify the extra expenditure of time which they demand. A modification of the Woody scales has displaced the A and B Series from very general use. The first series, remains, however, as the most useful scale to be found among current computation measurements for the detailed analysis of difficulties.

McCall has formed a scale from the four Series A scales. This edition is known as the Woody-McCall Mixed Fundamentals Test and consists of examples from the original scales arranged in order of difficulty but without regard to the operation they represent. This latter scale can be used in all grades from the second to the eighth, and grade norms are provided both for the grades as a whole and for low and high sections. The standardization is based upon a very large number of cases and fulfills all requirements of accuracy. A correction that is to be added to the grade norms permits the adaptation of these norms to any month of the school year, thus overcoming one of the more important objections to the usual type of norm. Four forms have been constructed, forms I and III consisting of thirty-five examples and forms II and IV of thirty-four. Examples taken at random from the Mixed-Fundamental Scale illustrate its form and content as well as one of the objections that may be voiced against it.

Examples From the Woody-McCall Mixed Fundamental Scale

FORM I

(1) Add	(2)	(6) Subtract	(16) Multiply	(23)
2		13	5096	
3	$2 \times 3 =$	8	6	$248 \div 7 =$

(24) Subtract	(27)	(30) Multiply	(34) Multiply
27		$987\frac{3}{4}$	$.0963\frac{1}{8}$
$12\frac{5}{8}$	$\frac{7}{8}$ of 624 =	25	.084

(35)

$25.091 + 100.4 + 25 + 98.28 + 19.3614 =$

The time allowance of twenty minutes is sufficient to afford all pupils an opportunity of solving all the examples that fall

within their ability. The score is the number of examples correctly answered. The survey function of this scale, and of this type of scale, is effected by means of the total score which represents ability in general computation. As it is inherently impractical in such measurements as these, no measure of speed is secured. Monroe has claimed that arithmetic possesses two dimensions, speed and accuracy, and that these scales involve accuracy only. It appears, however, that arithmetical ability is tri-dimensional, possessing the additional characteristic of power. Power is measured on scales alone, speed on tests alone.

As in all computation measurements that have been examined, the diagnostic function of these scales is an issue of supreme importance in educational measurements. It has been pointed out frequently that tests yield information regarding rate of work while neglecting methods and difficulties. Even under the most favorable circumstances a very limited number of types can be included without making the test so long that it could not be employed. The diagnostic value of a scale is conditioned largely by the number of types that it includes. In the scales that are being discussed the number of examples is thirty-five, though more types are involved than this number reveals. However, this number is too small for detailed work and is not sufficient to insure the degree of completeness that is desirable. This is caused by the use of but one scale for all grades. A number of important types have necessarily been omitted. Its value is thereby decreased approximately in proportion to the number of types that have been excluded. Notwithstanding this deficiency, experience has demonstrated conclusively the value that may be derived from the Woody-McCall scale through diagnosis of errors and consequent remedial teaching. Sr. Katharine McCarthy's study of errors in computation demonstrates its value in this function.[13] By comparing the accomplishment and improvements made by several grades, including experimental and control groups, Sr. M. Kathleen has reported the large improvement made by classes where the teacher was furnished with precise information regarding the difficulties that pupils en-

[13] Sr. Katharine McCarthy, O.S.B.: "A Study of Achievement in the Fundamentals of Arithmetic with Diagnosis of Errors." *Catholic Educ. Rev.*, xxii, 1924, pp. 533–539.

countered and the causes of such difficulties.[14] This latter experiment reveals the need of such measurements and the advantages that may be obtained from them. But the diagnostic value of the scales is reduced through the necessity of using but one scale for all grades. It appears that two scales are necessary for detailed analyses, one for the second, third and fourth grades, another for the fifth, sixth, seventh and eighth grades. The shorter form and greater convenience of the Woody-McCall Scale renders it more generally useful for ordinary purposes than are the Woody Series A scales. The greater precision of the latter, however, makes them superior when finer distinctions are to be made and more exact determinations are necessary.

Despite the many and important advantages which the Woody-McCall scales embrace, several criticisms have been directed against them. The availability of but one scale for all grades has already been noted. The inclusion of grade norms alone is an undesirable feature in educational measurements when age norms have the merit of greater precision and of being free from certain important deficiencies that are inherent within the grade norm method.

The reliability of the Woody-McCall scales is comparatively low according to results that have been obtained by Woody[15] and by Clark and Vincent.[16] Woody's results indicate that these scales possess a satisfactory reliability for class scores only. This reduces their value to a certain extent as survey measurements. Substantially the same results have been secured by Clark and Vincent in a study of variability. In all the grades studied the median scores were practically identical for the four forms. The median differences of individual pupil's scores, however, were large in comparison with the inter-grade intervals of the norms. The average difference between successive forms for all pupils was 2.36 examples. Such variability is the product of many factors, several of which are essentially characteristic of human performance. Woody's results are somewhat difficult to

---

[14] Sr. M. Kathleen: "Some Results of Remedial Instruction Following the Use of Diagnostic Arithmetic Tests." *Catholic Educ. Rev.*, xxiii, 1925, December.

[15] Woody, Clifford: *loc. cit.*

[16] Clark, John R., and Vincent, E. Leona: "A Study of Variability in Arithmetic." *J. of Educ. Psy.*, xvi, 1925, pp. 267–274.

explain when the reliability of the Woody-McCall scales is compared with that of other types of computation measurements. There is a possibility that the mixed-fundamental arrangement itself contributes to unreliability in various ways.

Owing to the fact that these scales are of the mixed-fundamental type it is necessary to preface many examples with a word or sign signifying the operation to be performed. In most examples McCall has employed the single word method. A reading element is thereby introduced into the scales which constitutes a disturbing factor and contributes to a reduction of their validity. Pupils' performances are lowered in a large number of cases through this arrangement. The amount of this reduction will vary according to the reading ability of the pupils. Lessenger reports that the loss in score due to the reading disability of pupils on a scale similar in form to the Woody-McCall amounted to approximately six months of achievement age.[17] This average represents a serious error that reduces the general value and particularly the validity of such forms of scales as these. Wilson's study of the influence of the reading factor led to similar results. The latter maintains that the arrangement and form of the scale are unsatisfactory. This criticism of the Woody-McCall scales is most important. It may account in part for the low correlations that have been obtained between this scale and other computation measurements, as well as for the low coefficients of reliability obtained by Woody, and the measures of variability secured by Clark and Vincent.

The reading problem involved has led to a more general criticism of the form of the scale. Wilson states that such an assortment of examples is likely to cause confusion. This criticism, however, appears to overlook certain fundamental issues involved in the normal functioning of arithmetical ability. Both Lessenger[18] and McCall[19] have called attention to the fact that this type of arrangement more nearly approaches the form in which problems are encountered in actual practice than do the more artificial forms adopted by Monroe and Courtis in their

[17] Lessenger, W. E.: "Reading Difficulties in Arithmetical Computation." *J. of Educ. Research,* xi, 1925, pp. 287–291.

[18] Wilson, W. K.: "The Woody-McCall Arithmetic Test as a Test in Reading. *Educ. Res. Bulletin,* iii, 1924, p. 329.

[19] McCall, William A.: *loc. cit.*

tests. The mixed fundamental form demands constant adjustments of the pupil to different processes. This adjustment is demanded by computation as it operates under ordinary conditions. If the conditions under which measurements are given should approximate those under which the ability is expected to function, this arrangement of the examples is easily justified. The observed fact that pupils encounter reading difficulties in these scales is probably evidence of the artificial conditions under which arithmetic is taught. If a pupil becomes confused by the form of this type of scale, it is probable that the confusion is but an expression of failure to master the essentials of computation. A distinction should be made between the single word direction method and the mixed form of the examples. That the single word direction method is a source of error cannot be denied. But the arrangement of the scale is not to be condemned because of the confusion that is sometimes involved, which merely reflects the methods of teaching whereby examples are grouped according to the processes. This grouping is highly artificial from many standpoints and while necessary in some respects should not engender a mental set that becomes very difficult to overcome. Judged by the standard of the ordinary usages of computation in after life, the arrangement employed by McCall is much less artificial than that involved in ordinary tests.

It is believed that this general type of computation measurement is the most satisfactory for general purposes. It lacks the disadvantages common to all speed tests. It possesses both diagnostic and survey values. The deficiencies that are present in the Woody-McCall scale do not constitute fatal objections either to this particular scale or to the type of measurement to which it belongs. Undoubtedly, several of the undesirable features might be overcome. The results that have been obtained from its use attest the value that it possesses.

The computation exercises of the Stanford Achievement Test are similar in all respects to the Woody-McCall scales. Two scales have been devised and included in the Stanford. The first includes twenty-five examples and is intended for the second and third grades. The advanced exercise contains twenty-two examples in addition to the twenty-five of the primary, or a total of forty-seven in all. Including as it does twelve more

examples than the Woody-McCall scale, a greater diversity of types is possible. The educational importance of some of the types may be questioned. The score is the number of examples correct. This may be translated into an age score by means of tables. It is, therefore, the first scale considered in which age norms are provided. Two forms of both primary and advanced exercises have been devised.

Some of the criticisms mentioned in connection with the previous scales apply with equal force to the Stanford. Of these, the principal one is the influence of the reading factor. As already mentioned, Lessenger's results indicate that inability to read the instruction word resulted in a loss of about six months of achievement age as an average. No information is available regarding the reliability of these scales other than that contained in the Manual accompanying the tests. The coefficients reported therein range from 0.76 to 0.92 for age groups from seven to fifteen years, inclusively.[20] All but one of these coefficients are above 0.83 and therefore very much higher than those secured from the Woody-McCall. Probable errors are not reported but, judging from the numbers that were evidently employed, they would be low. From these measures of reliability it may be inferred that unreliability is not characteristic of the type of measurement to which both the Woody-McCall and the Stanford belong but rather that it is due to some conditions or causes inherent in the former alone.

A desirable feature of the Stanford computation exercises is the presence of two exercises for the lower and upper grades respectively. The advanced exercise includes the primary. Had this arrangement been changed and the number of primary examples been reduced, it would have been possible to include a greater range of the more difficult types in the advanced exercise.

In several respects the Stanford exercises are superior to the Woody-McCall scales. The inclusion of a larger number of examples, the inclusion of age as well as grade norms, and the division of the scale into two exercises, are features that recommend its selection. Evaluated in the light of the principles previously discussed the Stanford is probably the most useful of all cur-

[20] Kelly, T. L., Ruch, G. M., and Terman, L. M.: "Stanford Achievement Test, Manual of Directions." The World Book Co., 1923, p. 63.

rent computation tests and scales. Its survey and diagnostic values combine to render it the superior of any test or scale that fulfills but one of these purposes alone. Its most apparent deficiency is the effect which reading disability has upon the scores.

The final scale which may be described here is the Pittsburgh Arithmetic Scale.[21] It has been included in the combination survey and diagnostic class, although in the latter respect it suffers from comparison with the Stanford and the Woody-McCall. It is constructed on a modified scale plan, having four examples at each level of difficulty. The four examples represent the four fundamental operations of computation. Three forms of the scale have been prepared and a preliminary exercise prefaces the actual scale. The employment of a preliminary exercise is a distinct advantage. Norms are given both for grades and for ages but the age scores are relatively crude since they are based upon six months intervals. Norms are also available for the number of examples attempted.

The agreement of the examples with educational objectives may be challenged. The method by which the scale was constructed necessarily involved the sacrifice of analytical value to statistical requirements. In order to keep pace with the rapid increase of the difficulty of the division examples, it was necessary to include problems in addition and subtraction that are out of harmony with the first principle of measurements. The following problems are particularly noteworthy examples of this deficiency:

(33)	(34)	(37)
Add	Subtract	Add
129		
666		54628
538		69835
729		76547
864		77698
979		48896
828	967382796	90388
763	78493887	94629

(38)	(39)	(40)
Subtract	Multiply	Divide
1370847090	87096	
750958089	809	85)695322550

[21] Guy J. Freeman: "The Pittsburgh Arithmetic Scales." University of Pittsburgh, 1923, p. 51.

No examples in fractions or decimals have been included. In the construction of this scale statistical principles of questionable necessity appear to have dominated with the resulting sacrifice of those examples that would have increased the diagnostic usefulness of the scale. Unless a separate test or scale of fractions be employed, computation abilities are not adequately tested. The diagnostic value of this scale appears to be small. In this respect it is analogous to the Courtis tests although speed is not a factor in the Pittsburgh scales. Merely to ascertain the number of examples missed in each of the four fundamentals is not diagnosis in the true sense of the word. Neither is it a diagnosis that assists a teacher to counteract definite sources of errors. To know that a child fails in more division examples than of any other type reveals nothing as to the cause of his failure. An insufficiency of types is a serious limitation of this scale.

This survey of the present status of computation tests and scales reveals the broad differences that exist among them in all respects. If the premises upon which this survey has been based are correct, the Stanford computation exercise represents the most useful scale of all. It fulfills the fundamental requirements of high reliability, diagnostic value, objectivity, and standardization. Its deficiency in validity through the necessary reading ability that it demands is the most important and most serious objection that can be raised against it. The class of measurements that it represents possesses undoubted superiority to tests that are either solely survey or solely tests of speed.

CHAPTER III

Methods of Test and Scale Construction

Not only do current tests and scales in computation differ in their form and content, but they also vary widely in the methods by which they have been constructed. As far as is known, no scale has been based solely upon an analysis of the abilities involved in computation generally. In contrast to the analytical method is the statistical method which has predominated in the formation of scales. There are indications that this latter method is not universally applicable when diagnostic value is the principal consideration. The involved and intricate methods pursued by Buckingham in the construction of a spelling scale, by Trabue in the formulation of his language scales, and by Woody in the field of computation scales, do not appear to be justified on the basis of results. The actual benefits that accrue to teachers from the employment of such highly refined scales are difficult to define and often impossible to discover. The primary qualification of an educational measurement of any description is the value it manifests in the uses for which it is designed. If scales permit the definition and localization of causes of errors and thus contribute to the effectiveness of teaching, it matters but little whether they have been constructed with or without the elaborate and involved methods of educational statistics. The usefulness of a test or scale seems to depend less on the methods by which it has been constructed than upon the content that has been selected. No amount of statistical manipulation will compensate for few or poorly chosen examples. The essential requirement in the construction of measurements, therefore, is not statistical finesse but general usefulness in the everyday problems of teaching. The varied methods that have been employed require an examination into their utility and their limitations. It is not intended to examine any particular procedures nor to describe specific statistical methods but to determine the general usefulness and necessity of such methods as have been applied in the construction of scales already examined.

As has been seen, computation tests contain a number of examples that are presumably equal in difficulty. A more precise definition of current tests would describe the examples that they

contain as being similar in appearance rather than equal in difficulty. Wide fluctuations and variations are to be found in the relative difficulty of different combinations. Thus, the percentages of pupils in all grades from the third to the eighth, inclusively, who failed on the following addition combinations were as follows:[22]

Example	*Per cent of errors*	*Example*	*Per cent of errors*
1 2	5.2	0 1	4.7
4 2	5.2	9 6	11.5
6 6	3.5	3 0	4.3
1 3	2.1	6 9	13.1
9 5	9.0	2 4	4.4
7 7	2.6	1 5	3.4

The above examples were taken from one of the tests described in the previous chapter. It is very evident that their difficulty is not uniform. This, however, does not constitute a criticism of such tests but rather an indication that the fundamental combinations are very unequal in difficulty and that no tests can be constructed that will include only examples that are equal. Difficulty will vary with type. The examples of the above group belong to two different types. In one the sum is less than ten, while in the second class are those examples where the sum is greater than ten. From this evidence, and from many other sources, it is evident that the determination of the difficulty of examples must have only relative accuracy. In the construction of many tests and scales, however, elaborate statistical processes have been based upon results that are no more constant than these and many others.

In the construction of tests, examples have been grouped more

[22] Clapp, Frank L.: "The Number Combinations: Their Relative Difficulty and the Frequency of Their Appearance in Textbooks." University of Wisconsin, Bureau of Educational Research, Bulletin No. 2, 1924, p. 120.

on the basis of their external appearance than upon exact determinations of their difficulty. It is true that if difficulty is to be the basis for the grouping of examples, a multiplicity of tests will be necessary. Similarity of appearance appears to have been, however, the determinant in the tests that have been described. The construction of speed tests involves little more than a process of standardization. In the formation of his General Survey Tests in Arithmetic Monroe has resorted to a system of weightings to obtain the total score.[23] These weights were adopted arbitrarily in order to apportion equally the contribution of each of the tests to the total score. The problem of weights in educational tests has not as yet been solved. It is obvious, however, that arbitrary procedures both in the selection of the examples and in the weightings of the tests have played an important part in the construction of computation measurements.

The methods of constructing scales are more involved and even more arbitrary. The necessity for many of the procedures is not easy to discern. Most of these methods have begun by ignoring the types of examples and have terminated in minute measures of difficulty. This is particularly pertinent to true scales where constant increments in difficulty have been the main determinants of the content of such scales. The most noteworthy example of true scales is found in the Woody Series B. These have already been examined. Their reliability is low and their diagnostic value inferior to the scales of Series A. Yet they have been constructed in a most precise manner and excel in statistical technique. The precision with which they have been built, however, has not contributed to their general value and the statistical complications appear to have been in vain. This naturally prompts the question of the necessity of such procedures in any case where tests and measurements are designed for classroom use. The methods employed by Woody in the case of both Series A and of Series B may be examined in some detail to ascertain whether they are necessary or not.

Woody selected a number of examples in the four fundamentals and formed tentative scales of these. The percentage of pupils who answered each example correctly was found for each grade. Only those examples were retained that were solved

[23] Monroe, W. S.: "The Illinois Examination," p. 19.

by an increasing percentage of pupils in successive grades. Gaps that appeared in scales were filled in with new examples. The difficulty of each example was then expressed in terms of the probable error for each grade. A normal distribution was assumed in this process, an assumption that was essentially valid. To determine the difficulty of the examples in terms of all grades combined, a zero point was necessary. The location of the zero point involved the finding of the intergrade intervals. Three methods have been ordinarily employed for this purpose. Woody used all three, assigning weights to the results obtained from the different methods. The weighted average of the three determinations yielded the intergrade interval from which in turn was found the zero point. The difficulty of all examples was then referred to this point and the location of each given in terms of the linear distance from zero. In the formation of the scales of Series B, only those examples were selected that were separated from one another by a constant amount. In the Series A scales the increments were not equal or constant.

The value of the above methods must arise from one or both of two sources. In the first instance, it is necessary in true scales. It has been shown, however, that scales resulting from the application of such methods possess but little value for ordinary purposes. The pragmatic test, therefore, does not justify or at least reveal the necessity of such an elaborate technique. The second of the possible values must arise through the greater accuracy of the scores. Woody has employed two methods of scoring. The scores obtained by the first method suggested by Woody were derived from the scale values of the examples which the pupils had answered correctly. According to the second method the score was simply the number of examples correct. The accuracy of the second method is sufficient to justify its use and, in fact, the first method has been displaced by the greater convenience of the latter. Apparently, therefore, the exact determinations of the difficulty of the examples has no weight in the final score. The diagnostic value of the scales of Series A is unaffected by the methods of scoring or the difficulty of the examples. The need of such refined measurements of difficulty, therefore, cannot arise from the contributions which such methods make to accuracy of survey scores or to diagnostic value.

The question of the accuracy of the zero point may also be

raised. Changes in the curriculum, differences in the group studied, time of year, these and many other factors influence the location of the zero. The structure built upon such framework must obviously be weak. Its accuracy is more apparent than real. If a change should be effected in the location of zero through any of the numerous causes that affect it, the whole elaborate system of scoring is thrown out of harmony.

The ultimate criterion of the form, content and methods of construction of measurements is the value of such tests and scales in teaching. If such methods of construction do not contribute to the general usefulness of the scale they are superfluous from this point of view. In the case of the Woody scales there is no evidence that any contribution to general utility is rendered through the methods that the author of such scales has employed. It would appear that as reliable and as useful scales could be constructed without any such methods or at least with much simpler ones.

Monroe has recently stated that ". . . some recent studies of the scoring of scaled performance tests indicate that serious errors would not be introduced if the procedure were greatly simplified."[24] The methods suggested by Monroe are the determinations of difficulty on the basis of combined grades or on the basis of one grade alone, preferably a mid-grade. Results obtained by Clapp appear to justify the latter procedure for the fundamentals of arithmetic.[25] This writer found that the number combinations tend to maintain the same relative difficulty in all grades. The zero may then be referred to the distribution within a single grade and would probably possess as much stability as when computed by the longer and more intricate method. Either of the procedures suggested by Monroe would greatly simplify the process of scale construction. Unless true scales are to be formed, the zero point is not essential; for the number of correct examples is as satisfactory a score as one derived from the scale value of the examples. Judged, therefore, by either the standard of necessity or of accuracy, the highly refined statistical methods pursued by Woody in the construction of his computation scales do not appear essential to a valid and reliable scale.

---

[24] Monroe, W. S.: "An Introduction to the Theory of Educational Measurements." New York, 1923, p. 91.

[25] Clapp, Frank L.: *Loc cit.*

The construction of the Pittsburgh Arithmetic Scales has been based on essentially similar methods but exaggerates the importance of scaling and sacrifices to it practically all diagnostic value. A large number of preliminary examples were employed and from these were selected those that were incorporated into the final scales. The extent of the selection may be judged from the fact that the preliminary scales included 263 examples while the final scale has but forty. A method suggested by McCall was adopted. The difficulty of each of the 263 examples was found from the percentages of twelve-year-old children who answered these examples correctly. From these percentages were found the percentile scores of each example according to Rugg's tables. Ten examples were then chosen from each of the fundamental processes that were equally spaced on the PS scale. By this method there is one example from each process at each interval on the scale. The only criterion, therefore, that was employed in the selection of the examples was their PS or scale location. No attention whatever was paid to the diagnostic value of the examples. The final scale represents a true scale that has been constructed in accordance with rigid statistical principles to the utter neglect of all other considerations. General usefulness has been subordinated to statistical precision. Diagnostic values had no influence whatever in the selection of the examples. The Pittsburgh scale, therefore, typifies the results that may be expected when statistical behavior and not diagnostic value is the fundamental factor involved. The sole resulting advantage is the presumed equality of the three scales. Guy states the question at issue: "Statistically, Forms A, B, and C are constructed on substantially equal PS values. Is a score, then, of 27 Right on Scale A equal to a score of 27 Right on Scale B or C? *Statistically, yes. Practically, we do not know.* The answer to this question cannot be determined until norms and distribution are established for Forms B and C as have already been done for Form A."[26] This admission of the absence of any degree of finality to the PS method and the necessity of experimental rather than statistical determinations of the equality of the scales appears to be a fatal objection to the method applied. If determination remain necessary, the preceding method is nothing but superfluous; for scaling and equalizing of the forms could be more readily accomplished through the T score method

[26] Guy, J. Freeman: *loc. cit.*, p. 25. (Italics mine.)

as has been done by McCall for the Thorndike-McCall reading scales. The advantages of this technique are few if any. The limitations of the final scale, on the other hand, are manifold and most of them may be traced directly to the methods that have been applied. A very different problem from the scaling of examples is the scaling of scores. Scaled scores secure most of the advantages sought through the scaling of the examples themselves. The methods ordinarily used are comparatively simple. Of these methods the T score technique of McCall is probably the best known.[27] The principal limitation of McCall's method lies in the narrow scope of such scores. Scales intended for the early grades or for high school subjects do not fulfill the necessary conditions of such scaling. The method is the most simple one available for the equating of forms which in themselves may be of unequal difficulty. Had this method been substituted in the derivation of the Pittsburgh scales, the problem of inequality would not have arisen.

It is the opinion of the present writer that many of the methods employed in the construction of scales are entirely superfluous while others do not compensate by their results for the labor which they entail. The supposed advantages accruing from such methods are neither real nor apparent. In neither the Woody scales nor in the Pittsburgh scales is any greater diagnostic value available from the elaborate procedure alone. The essentials of scale construction can readily be secured by other means. If examples are arranged in order of difficulty without any large unfilled intervals and the resulting scores are scaled by the T score method, all the merits of the more elaborate methods are secured without their limitations. The adoption of unmodified statistical methods almost inevitably leads to the exclusion of important types of examples. General usefulness becomes subordinated to abstract principles of doubtful validity. Diagnostic value is sacrificed with no apparent return. The aim in the construction of computation scales should be to make scales that will possess the maximum survey and diagnostic values, that will prove most useful and accurate, not scales that will attain problematical exactitude at the cost of essential characteristics and properties.

---

[27] McCall, William A.: "How to Measure in Education." New York, 1921, p. 416.

CHAPTER IV

The Diagnostic Computation Scale

The principles that were described in Chapter I have been employed in the construction of scales for the measurement of ability in computation. The survey of the present status of such measurements has revealed certain requirements underlying such tests and scales and the desirable characteristics that computation scales should manifest. The immediate aims in the construction of the new scales may be stated as follows:

1. The advantages which scales possess over speed tests dictate that computation measurements generally should be of this type. The value of speed tests is not to be overlooked but the more fundamental type is the scale.

2. In order that the maximum benefit may be derived from a minimum of time and effort, the mixed-fundamental type has been selected. This type has the additional merit of conforming more closely than do others to the situations in which computation ability normally functions.

3. The validity of such scales should be as high as it is possible to render it. For this reason the single word direction method has been discarded in favor of signs to designate the process involved in the example.

4. The diagnostic value of these scales has been made as great as possible. As many different types of examples have been included as the limitations of length would permit. This requirement has been emphasized throughout the construction of these scales and it is intended that their value in this respect should excell that of other scales of a similar type already discussed.

5. In order to promote their diagnostic value and to include as many examples as possible, two scales have been formed. Part I includes the more fudamental types and is intended for the second, third and fourth grades, while Part II contains fewer easy examples and many additional difficult ones and is intended for the fifth, sixth, seventh and eighth grades. The formation of two parts permits the exclusion of many examples for the upper grades that children in such grades answer with practically perfect accuracy.

6. Several forms of equivalent difficulty have been constructed to permit re-testing.

7. Specific directions for the administration and scoring have been included in order to render objectivity as high as possible.

8. In the selection of types of examples, educational objectives in arithmetic have constituted the main determinants in the inclusion or exclusion of certain types. No example has been included on account of its scale value alone. It is believed that all important types have been included.

9. As far as may be controlled by the construction of the scales, reliability has been an important consideration. The inability to control many of the sources of unreliability, however, prevents adequate supervision of the observance of this requirement.

10. Norms have been based upon a representative distribution, including in all approximately 10,000 scores. Both age and grade norms have been found. For Part II, T scores have also been derived.

11. The time limits have been so arranged as to permit practically all pupils to answer all examples within their power.

12. All degrees of ability are represented through the inclusion of both very easy and comparatively difficult examples. Examples should not, however, be included merely because they are difficult. The presence of perfect scores on a computation scale is not an evidence of weakness unless important types have been omitted.

Using these aims as definitions of the form and content of the scales, the processes involved in their construction may be described in order to evaluate the conformity of the products with the requirements governing their construction and function. In order to insure high diagnostic value a large number of types have been included. Examples of these types formed the preliminary scales that were used for the determination of the difficulty of the items and for the standardization. The large number of types rendered it imperative that two scales be formed, one for the early and one for the more advanced grades. This method possesses the distinct advantage of permitting the inclusion of many examples of low difficulty for younger children and of high difficulty for the pupils of the upper grades. A number of types are common to both scales. The point of divi-

sion was the lower limit of the fifth grade. This basis was employed on account of the general introduction of fractions and decimals in this grade. Two preliminary scales were accordingly formed, Part I being for grades two, three and four; while

Diagnostic Computation Scale

Part I, Form A

(1)	(2)	(3)	(4)	(5)	(6)
6 +1	4 6 1 2 5	8 −4	3×6=	7−3=	9 +6
(7)	(8)	(9)	(10)	(11)	(12)
54 63 28	37 −15	14 −9	8+9=	49 −20	2)0
(13)	(14)	(15)	(16)	(17)	(18)
25 −2	34 +5	18 +7	0 ×8	3 ×7	32+17=
(19)	(20)	(21)	(22)	(23)	(24)
5+3=	1)1	53 ×2	7)14	9 27 632 45 83	5)30
(25)	(26)	(27)	(28)	(29)	(30)
154 −81	1234 5971 2038 7156	53 −8	32 −24	24 ×7	82 −57
(31)	(32)	(33)	(34)	(35)	(36)
406 ×3	6572 ×4	1248 −759	247 ×9	7)36	6)756
(37)	(38)	(39)	(40)	(41)	(42)
3000 −467	163 ×40	36)396	8)727	54)4482	7236 ×28
(43)		(44)	(45)		(46)
23)13938		47)1139	6283 ×206		462)24490

Diagnostic Computation Scale

Part II, Form A

(1) 21×8

(2) $23 - 12$

(3) $53 - 20$

(4) $43 + 84 + 76$

(5) $74 - 26$

(6) $57 - 48$

(7) 26×4

(8) $32 - 7$

(9) 304×6

(10) $142 - 61$

(11) $\frac{2}{3}$ of $18 =$

(12) $\$4.25 + 3.16 + 2.11 + .93 + 1.05$

(13) 163×9

(14) 4389×6

(15) $64 + 382 + 53 + 68 + 45$

(16) $8\overline{)35}$

(17) $43\overline{)516}$

(18) $2434 + 7863 + 4289 + 1375$

(19) $1.26 + .95$

(20) 128×70

(21) $\$5.00 - .38 =$

(22) $2654 - 826$

(23) 0×8

(24) $8.624 - 3.985$

(25) $\$20.00 - \$4.85 =$

(26) $\frac{1}{4} \times 15 =$

(27) $\frac{1}{3} + \frac{1}{3} =$

(28) $54\overline{)21924}$

(29) $283 \times .08$

(30) $3\frac{3}{4} - 1 =$

(31) $\frac{1}{3} \times {}^{3}/_{5} =$

(32) 6251×37

(33) $3\overline{)63.54}$

(34) $32 + 16\frac{3}{4}$

(35) $12 \div {}^{4}/_{5} =$

(36) $7\overline{)492}$

(37) $62.1 \times .42$

(38) 1 ft. 8 in.
5 ft. 6 in.
3 ft. 3 in.

(39) $27 - 6^{2}/_{5} =$

(40) 20% of $\$600 =$

(41) $6.4 - 1.231 =$

(42) ${}^{1}/_{6} + \frac{1}{3} =$

(43) $\frac{2}{3} \div 5 =$

(44) $27 \times 3\frac{2}{3}$

(45) $\frac{1}{4} + {}^{5}/_{6} =$

(46) $3\frac{1}{2} \div 2\frac{1}{3} =$

(47) $\frac{7}{8} - \frac{2}{3} =$

(48) $52\overline{)1884}$

(49) ${}^{5}/_{6} \div \frac{3}{4} =$

(50) $3\frac{1}{4} - 1\frac{1}{2} =$

(51) $3.6\overline{)7.812}$

(52) $264 \times 12\frac{1}{2}$

(53) $3\frac{3}{4} \times 2^{1}/_{6} =$

(54) $6.5 - 4.0045 =$

(55) $16.45 \div 1\frac{1}{4} =$

(56) $.16\overline{).9}$

(57) $68\frac{1}{3} \times 14$

(58) $.24\overline{)1272}$

Part II was employed in the remaining grades. Each part had three forms, designated as A, B, and C, which were identical in the number and types of examples included but differed in the content of the types. The three forms of Part I each consisted of forty-six examples, while those of Part II embraced fifty-eight each. Form A of the revised scales is reproduced in reduced size on the previous pages. The only differences existing between the preliminary scales and the revised forms are in the sequence of the examples and in a few minor details. The number of examples included is considerably greater than in any other scales of this type. The primary exercise of the Stanford Achievement Test includes twenty-five examples while Part I of the new scales has forty-six. The advanced exercise of the Stanford has forty-seven examples in contrast with fifty-eight of the present scale. It is thus seen that a much greater variety of examples is made possible in this way.

The preliminary scales were printed in a four-page folder, on the cover of which were blanks for the usual information such as the pupil's name, grade, age, school, date, etc., and a preliminary exercise. The value of preliminary exercises in educational measurements has always been obvious from experience with tests in classes. This is particularly true when the children are unaccustomed to such measurements and the conditions under which they are given. The preliminary exercise serves to accustom the pupils to the directions and to equalize any extraneous factors that might prove disturbing. A pamphlet containing instructions for the administration of the scales accompanied them. These instructions were carefully designed to occupy a minimum of time while making explicit the nature of what was required of the pupils. The time limits were sufficient to permit practically all pupils to reach the limit of their ability. It is unlikely that the scores would have been materially affected had this limit been increased. Many children, of course, completed all the examples that fell within their grasp before the expiration of this period. An initial trial of the preliminary forms had provided a rough indication of the difficulty of the examples and prevented the location of any easy examples toward the end of the scales.

A particular difference between these scales and the Woody-McCall and the Stanford computation exercises is the substitu-

tion of signs for words to designate the operation. The objections to the word direction method are so serious that it was deemed necessary to avoid it. The use of signs does overcome the reading difficulty. All pupils learn the meaning of signs in connection with computation. It is, therefore, believed that extraneous abilities are not introduced through such a substitution. As will be seen later many mistakes were made in the interpretation of the signs but this does not constitute an objection to their use. Probably the sign method is not altogether satisfactory but it is a lesser evil than the single word method.

The preliminary forms of the scales were given to approximately 10,000 pupils, distributed as follows through the grades:

Table I—Distribution of Pupils for Preliminary Forms

	Grades						
	II	III	IV	V	VI	VII	VIII
Part I.							
Form A....	501	471	523	268			
Form B....	227	316	527	293			
Form C....	464	677	638				
Total....	1192	1464	1688	561			
Part II.							
Form A....		162	138	402	446	407	345
Form B....				457	441	579	413
Form C....				499	480	443	381
Total....		162	138	1358	1367	1429	1139

Total for all forms and all grades: 10,498.

The number of cases involved together with evidence from the normality of the distributions indicate that a sufficient number of scores is involved to insure a satisfactory degree of accuracy and reliability. The preliminary scales were all given at the end of May and the beginning of June, 1925. An effort was made to have all scales given at approximately the same time to secure, as far as possible, uniformity of conditions. A wide selection of schools was also made in order that the resulting norms might be truly representative of computation ability throughout the grades. The extent to which children who took these tests had had experience with other educational measurements is unknown but probably this has had little or no effect

upon the measures of difficulty or the scores. Included within the manual of directions were instructions governing the scoring of the scales and the tabulation of the results. All processes have been verified and the errors reduced to a minimum as far as is known.

The difficulty of all examples was calculated for each grade separately. It was observed that Part I of the scales was entirely unsatisfactory for use in Grade V and the results obtained from its use in this grade have not been included in the final results. The difficulty of each example was first expressed in terms of the percentage of pupils for each grade who solved the example in question. On account of their length these tables have not been included within the present monograph. These initial determinations revealed a number of conclusions of importance in the construction of computation measurements. The examples tend to maintain their relative difficulty in all grades. This is an extension of the results obtained by Clapp whose study was confined to the number combinations. This observed fact also demonstrates the validity of the suggestion advanced by Monroe regarding the construction of scales. Presuming that the examples do maintain their relative position it would be sufficient to determine their difficulty for one grade alone instead of for each grade separately or for all grades combined. This procedure would greatly abbreviate most methods of scale construction. It would be necessary in such a method to employ results from a middle grade, such as the fifth, if the test or scale is to be used for all grades of the elementary school.

The method used in this study has been to ascertain the difficulty of the examples on the basis of all grades in which the scales are intended to be used. The difficulty is then expressed in terms of the percentages of correct answers. While it is customary, and sometimes necessary, to express difficulty in terms of a scale value, such as the standard deviation or the probable error, only the percentages are given in the accompanying tables. These percentages may readily be converted into terms of the probable error or sigma by means of tables of equivalents. In view of the greater significance to the majority of teachers of percentages, they alone have been reported. In the tables the difficulty of all examples is given on the basis of combined grades. The numbers of the examples refer to their order on the revised

scales and correspond to the numbers on the scales that were reproduced on previous pages.

A number of necessary changes were made in the final scales. Five examples have been added to Part II. The difficulty of these examples was determined independently of the preliminary scales. The accuracy of their location is sufficient for all ordinary purposes. No examples have been added to or subtracted from Part I. A few minor changes in other respects have also been made. These are so slight, however, as to have no influence on the scores. The spacing of the examples has been increased.

Table II—Percentages of Correct Answers

Part I, Forms A, B, C

Number of Pupils, 4325

Ex. No.	Per cent	Ex. No.	Per cent
1	90.3	24	63.6
2	87.2	25	62.3
3	86.8	26	60.8
4	86.7	27	60.7
5	85.5	28	60.3
6	84.0	29	59.0
7	83.5	30	58.3
8	81.5	31	55.6
9	79.8	32	45.2
10	77.5	33	45.0
11	77.0	34	44.0
12	77.0	35	43.4
13	75.6	36	34.0
14	75.6	37	32.8
15	74.6	38	29.4
16	74.6	39	25.5
17	74.1	40	25.1
18	73.6	41	20.2
19	72.2	42	20.0
20	69.8	43	19.3
21	68.5	44	17.1
22	67.2	45	15.3
23	63.8	46	7.6

The zero point of the scale has not been determined other than in connection with the scaling of the scores. The method of scoring that was adopted is the usual one—the number of correct examples. The use of this method does not necessitate the determination of the zero point and, as such a process would have contributed nothing to the general value of the scale, it has been

entirely omitted. The rearrangement of the examples in accordance with their difficulty and the inclusion of a few additional examples to represent certain important types constitute the only substantial changes made in the scales in the course of their revision.

The final or revised scales include Part I for the second, third and fourth grades, and Part II for the fifth, sixth, seventh and eighth grades. For each part there are three forms of substantially equal difficulty. They are printed in four-page leaflets with ample space for the recording of answers and for any computations that are necessary. The first page contains blanks for such information as may be required and a preliminary exercise. Practically no change has been made in the form of this cover.

Table III—Percentages of Correct Answers

Part II, Forms A, B, C

Number of Pupils, 5284

Ex. No.	Per cent	Ex. No.	Per cent
1	97.0	30	70.0
2	96.8	31	69.4
3	96.8	32	67.5
4	95.0	33	64.1
5	94.1	34	63.9
6	94.0	35	63.3
7	94.0	36	60.0
8	93.6	37	58.3
9	93.2	38	51.7
10	91.3	39	51.6
11	90.3	40	50.0
12	88.4	41	48.9
13	88.4	42	42.6
14	88.3	43	42.3
15	87.3	44	40.6
16	87.3	45	39.0
17	86.0	46	38.4
18	85.0	47	37.1
19	85.0	48	36.4
20	83.0	49	35.4
21	82.6	50	34.9
22	81.3	51	33.1
23	80.8	52	32.4
24	76.0	53	29.6
25	73.3	54	29.3
26	72.8	55	24.5
27	71.8	56	24.5
28	71.3	57	22.5
29	70.6	58	20.6

CHAPTER V

The Standardization of the Diagnostic Computation Scale

As has been seen in the preceding survey of computation measurements, various systems of scoring have been employed. Of these the most common is the point score method derived from the number of examples which have been answered correctly. Weightings may be added to this system in such instances as the Stanford Achievement Test where the computation exercise is but one of a series contributing to a general measure of accomplishment. Various derived scores are sometimes obtained from such point scores. The limitations of point scores and the consequent necessity for derived scores are generally recognized. In order to compare achievement in different subjects a common denominator is needed in terms of which accomplishment may be expressed. The comparison of achievement scores with those obtained on intelligence tests necessitates that both be in terms of a common unit. Point scores may be validly compared with point score norms only at certain times of the school year unless corrections are supplied. The general difficulty of interpreting such scores is one of their principal limitations, for the meaning attached to them differs for all tests and scales. Any test or scale, therefore, in the majority of subjects of the elementary school should have norms and scores in terms of both crude scores and of some generally accepted system of derived scores.

Of the scales and tests examined in this monograph, the Woody Series A alone presents an unusual system of scoring. The method devised for these scales is intricate and difficult to use. The number and percentage of pupils solving each example is ascertained. The problem that can be answered correctly by exactly 50 per cent of the class is used as the best measure of the ability of the class on these scales. The five problems that most closely approach this measure of difficulty are selected and the established value for each is found from tables supplied in the handbook of directions. By means of a second table the amount that must be added to or subtracted from these values is found. The average of the five resulting measures of difficulty represents the class score. This method assumes the prior processes of scaling the examples in terms of a zero point as has been already described. The whole procedure is lengthy and it is

doubtful if it possesses any substantial advantages over a much shorter and more readily understood method. On account of the omission of the scaling process in the present scales, the system employed by Woody cannot be applied.

Woody has also used the number of examples answered correctly as the score. The accuracy of this method is sufficient for all ordinary purposes. It is the most generally applied method for all educational measurements and possesses the distinct advantage of being easily understood and followed. Its most serious limitation is due to its failure to weight the credit for the solution of the more difficult examples. All problems have an equal value, despite their great inequalities of difficulty. To weight the more difficult examples, however, is a complicated procedure and the high correlation obtained between weighted and unweighted scores justifies the abbreviation of the procedure and the use of the shorter method. The method of scoring involved in The Diagnostic Computation Scales is the point score method, giving a score of one for each example that is solved. From these point scores various derived scores may be found. Among the more common derived scores are age scores, T scores and B scores. Age scores possess the merit of greater ease of interpretation, for the use of the other systems is not sufficiently widespread as yet for them to convey the same significance as that of age scores.

Both grade and age standardizations are available on the present scales. The grade norms were obtained from the median number of examples correct for all grades in which the scales are to be used. Corrections have not been added to these to adapt them to each month of the school year in view of the presence of age norms. The norms represent the average ability of the pupils in the grades designated for the end of the school year. The grade norms for both parts of the scales are contained in table IV.

Table IV—Grade Norms

	Grade							
	II	III	IV	V	VI	VII	VIII	No.
Part I....	18.8	26.3	35.4	.....	.....	.....	.....	4323
Part II...	.....	14.3	25.8	29.7	35.0	40.7	44.4	5402

The accuracy of these norms is judged sufficient from the number involved and the normality of their distribution. The large number of schools included within this study and their representative character contribute to the stability of the medians. Norms are included for the third and fourth grades for Part II, although it is not intended that this part of the scale should be used in these grades. However, they permit the evaluation of sub-average accomplishment in the fifth grade and under certain circumstances may be used, but Part I is better adapted to grades below the fifth.

Table V—Distribution of Scores, Part I, Combined Forms A, B and C

Number correct	II	III	IV	Number correct	II	III	IV
0	5	.....	1	25	37	63	35
1	4	1	.....	26	32	39	50
2	9	1	.....	27	30	57	31
3	12	8	3	28	35	58	53
4	13	7	.....	29	36	66	53
5	16	10	.....	30	27	67	56
6	20	8	1	31	29	61	46
7	27	12	4	32	19	53	54
8	39	14	2	33	15	62	67
9	34	22	1	34	9	47	86
10	45	22	2	35	8	54	82
11	30	24	7	36	5	44	68
12	43	25	6	37	2	39	84
13	40	31	6	38	2	29	92
14	53	30	5	39	1	23	89
15	50	25	7	40	.....	14	100
16	50	29	8	41	1	13	95
17	68	39	10	42	.....	2	111
18	47	46	12	43	.....	4	86
19	50	40	15	44	.....	4	66
20	59	40	14	45	.....	.....	51
21	56	45	17	46	.....	.....	26
22	40	58	29	.....	.....	.....	.....
23	39	53	21	N	1189	1446	1688
24	52	57	36	Md	18.8	26.33	35.35

An age standardization has been secured for both parts. The median score of all ages exclusive of the extremes was found for intervals of two months. Owing to the few scores obtained for the extreme ages, a certain amount of artificiality is necessarily

introduced through exterpolation. Children in the second grade who are under seven years of age cannot be considered as representative of all children of that age. The same condition is found at the upper limit of the scales. The medians obtained for each age group have been smoothed by the moving average method. This method was deemed superior to the other methods for this type of data. The age norms for Part I are presented in the accompanying table. The extension of the norms to the lower ages is necessarily arbitrary. The high scores obtained by the youngest groups are due to two factors. The children of the lower ages included within this study are undoubtedly superior children whose scores are not a correct estimate of the ability of all children of the same age. Furthermore, the inclusion in the scale of a large number of very easy examples has contributed to high medians for the primary age groups. The value of such examples from the diagnostic and survey viewpoints is so great that they have been retained despite the effects

Table VI—Age Scores

Part I

Number correct	Age score	Number correct	Age score
1	6–4	24	8–8
2	6–5	25	8–10
3	6–6	26	9–0
4	6–7	27	9–2
5	6–8	28	9–4
6	6–9	29	9–6
7	6–10	30	9–8
8	6–11	31	9–10
9	7–0	32	10–0
10	7–1	33	10–2
11	7–2	34	10–4
12	7–3	35	10–6
13	7–4	36	10–8
14	7–5	37	10–10
15	7–6	38	11–0
16	7–8	39	11–3
17	7–9	40	11–6
18	7–10	41	11–9
19	7–11	42	12–0
20	8–0	43	12–3
21	8–2	44	12–6
22	8–4	45	12–9
23	8–6	46	13–0

which they have upon the standardization. From the absence of more than a very few zero scores it appears that this scale is capable of estimating all degrees of computation ability and certainly the lower degrees.

Age scores are the only derived scores employed in Part I. The T score method used with Part II cannot be applied without considerable modification to the earlier age group. It has been suggested that a nine- or ten-year group be substituted for the twelve-year group of McCall's method, but up to the present this suggestion has not been carried out.

Two systems of derived scores have been employed for Part II. These have the additional value of checking one another and establishing the accuracy of the entire standardization. The first derived scores are T scores and were determined according to the usual method suggested by McCall.[27] The validity of the resulting scores depends almost entirely upon the representative character of the distribution from which they are obtained. The number of children between the ages of twelve and thirteen in all grades was 1,165, which is regarded as sufficient for the purpose. The median score of this group was 36.68 examples with a standard deviation of 9.75. A score of 50T accordingly is equal to 36.68 correct examples and T equals .975 or approximately one example. The remaining equivalents for all scores have been found and are included in the table of norms for Part II.

On several tests and scales both age scores and T scores are included. The methods and meaning of these scores lead to certain conclusions and serve as a check of the accuracy of any standardization that employs both systems. By the definition a T score of 50 is always equal to an achievement age of 12 years 6 months or 150 months. It may be assumed that variability of chronological age is a constant whenever the group studied is sufficiently large to give a random sampling. The validity of this assumption may be demonstrated from empirical results. It necessarily follows from this assumption that a T score is always equal to a certain age and to no other age. Hence all T scores will have definite equivalent subject or achievement ages irrespective of the scale involved. If this assumption is

[27] McCall, William A.: "How to Measure in Education." New York, 1921.

correct any T score should have the same corresponding age score on such scales as the Thorndike-McCall Reading or the Morrison-McCall Spelling Scales. An examination of the tables for the converting of scores into various derived scores supplied with these scales shows that in general a T score is equal to one and only one achievement age irrespective of the scale. The discrepancies appear only at the extreme ages where various factors operate to invalidate the results. It follows also from this hypothesis that the age standardization of any scale can be accomplished merely through the determination of T scores and the adoption of either the Morrison-McCall table of equivalents or that contained in the Thorndike-McCall scale. Of the two, the latter shows a straight line increase, while certain irregu-

Table VII—Age Scores and T Scores

Part II

Number correct	Age score	T score	Number correct	Age score	T score
1	6–3	13	30	10–10	43
2	6–4	14	31	11–1	44
3	6–5	15	32	11–3	45
4	6–6	16	33	11–6	46
5	6–8	17	34	11–9	47
6	6–10	18	35	12–0	48
7	7–0	19	36	12–3	49
8	7–1	20	37	12–6	50
9	7–2	21	38	12–8	51
10	7–3	22	39	12–11	52
11	7–5	24	40	13–3	53
12	7–8	25	41	13–6	54
13	7–9	26	42	13–9	55
14	7–10	27	43	14–0	56
15	8–0	28	44	14–4	58
16	8–2	29	45	14–8	59
17	8–3	30	46	14–11	60
18	8–4	31	47	15–1	61
19	8–6	32	48	15–4	62
20	8–8	33	49	15–6	63
21	8–10	34	50	15–9	64
22	9–0	35	51	15–11	65
23	9–3	36	52	16–1	66
24	9–6	37	53	16–3	67
25	9–8	38	54	16–5	68
26	9–10	39	55	16–7	69
27	10–1	40	56	16–9	70
28	10–4	41	57	16–11	71
29	10–7	42	58	17–1	72

Table VIII—Distribution of Scores, Part II, Combined Forms A, B and C

No. correct	III	IV	V	VI	VII	VIII	Total
0			2			1	3
1	2		2				4
2	2			1			3
3	1	1					2
4	4	2	1	1			8
5	2		2				4
6	6		5	1			12
7	6	3	9	1			19
8	10		5				15
9	5	1	7	1		1	15
10	6		8		1		15
11	13	1	6	4			24
12	8	4	8	2	1		23
13	10	2	13	7	2		34
14	16		15	10	1	1	43
15	12	2	11	6	1	1	33
16	12	3	17	5			37
17	16	7	23	8	1	3	58
18	6	4	20	17	9	1	57
19	8	3	35	12	5	2	65
20	4	7	38	19	4	2	74
21	6	5	27	26	10		74
22	5	5	43	24	8	5	90
23	1	6	38	29	13	5	92
24		8	53	31	9	9	110
25		6	56	33	12	10	117
26		12	55	42	27	4	140
27		12	69	34	15	6	136
28		13	65	50	26	13	167
29		5	66	44	26	14	155
30		9	67	47	33	18	174
31		7	60	64	36	11	178
32		3	65	54	37	21	180
33		1	55	60	33	16	165
34		4	55	56	36	15	166
35		1	55	57	50	23	186
36		0	51	70	51	31	203
37		1	27	61	46	22	157
38			34	62	52	45	193
39			40	49	60	30	179
40			34	51	65	49	199
41			23	38	57	56	174
42			15	45	66	47	173
43			15	43	58	45	161
44			15	30	69	66	180
45			8	36	65	71	180
46			6	19	57	58	140
47			8	29	57	61	155
48			8	26	49	67	150
49			5	23	41	58	127
50			1	21	38	78	138
51			3	16	34	41	94
52			4	9	28	38	79
53			2	7	10	22	41

larities are to be found in the increases of the Morrison-McCall scale.

Assuming the validity of this hypothesis on the basis of the nature of T scores and variability in age as well as the corroborative evidence from the above-mentioned scales, the accuracy of the age standardization for Part II may be judged. Inspection of the results both in tabular and graphic form confirms the hypothesis that has been advanced in this study. Certain irregularities at the earlier ages characterize the present computation scale. These irregularities are almost certainly due to the selective character of the scores for the earlier ages. The conformity of the norms on the Diagnostic Computation Scale with those of the other scales may be accepted as evidence of the accuracy of the present standardization. The T scores and age scores corresponding to the number of examples correct are given in Table VII.

As is always found, the amount of overlapping between successive grades is very great. While sometimes interpreted to signify low discriminative capacity, such overlappings of the distributions reveal the heterogeneous grouping of pupils as it exists in schools at the present time. This overlapping must be interpreted as the result of the present organization of classes rather than due to defects inherent within the scales themselves.

CHAPTER VI

The Diagnostic Value of the Scale

The uses which educational measurements fulfill may be classed as survey and as diagnostic. All other functions are but divisions of these. The survey or comparative values of a computation scale are indeed important. Information is required by teachers and supervisors regarding the level of attainment that a class has reached. But this level of accomplishment is conditioned by the habits they have formed and the skills they have developed. Disability requires analysis. It is not only important to know that a class is average, below average or above average, but it is more essential to determine the sources and causes of the errors that are made by the class as a whole or by individual pupils. Errors in computation are not merely errors but evidences of difficulties and incorrect methods. To know that a pupil is below the average dwindles in importance before such information as that which conveys knowledge of why he is retarded. Diagnostic measurements that assist the teacher to locate difficulties to the end that they may be overcome contribute more generously than any purely survey or speed test ever can. The diagnostic function of a computation scale is the principal one and its general value may be evaluated by the completeness with which difficulties may be analyzed through its use.

The majority of errors committed by a class group themselves into certain types which will vary for grades and for schools. If these types can be identified the process of overcoming them is greatly facilitated. A study of the results obtained from computation measurements entails no difficulties that teachers cannot master. If the types can be identified the process of remedial teaching will be defined and the results secured will be much greater than when no such diagnosis is made. These statements are verified by the results secured by Sr. M. Kathleen.[28] The progress made by a control group was com-

[28] Sr. M. Kathleen. *Loc. cit.*

pared with that accomplished by two other groups in which the children's difficulties had been analyzed and the teacher furnished with the necessary information upon which to base remedial teaching. In one of the experimental groups only class errors were considered whilst in the other a record was made of individual difficulties. The time devoted to arithmetic in all grades was the same and the groups were substantially equal both in initial ability and in general intelligence. The scales (Woody-McCall Mixed Fundamental) were repeated at the end of six weeks and the relative progress of the several groups was measured. The following summary of the results shows the great increase manifested by the experimental schools.

Table IX

School	Experimental		Control	
	I	II	I	II
Grade VI	25.9	29.5	26.0	27.7
Grade VII	27.5	33.6	28.2	30.2

GAINS

School	Experimental	Control
Grade VI	3.6	1.7
Grade VII	6.1	2.0

The results obtained in this study of the use of diagnostic tests are corroborated by others which, however, failed to make any comparisons with control grades. Employed for diagnostic purposes such scales as those that belong to the mixed fundamental type are capable of assisting the teacher in increasing the ability of the class to a greater degree than is possible without their aid. The methods that are necessary in such procedures are quite simple and much can be accomplished with a minimum expenditure of labor and money.

An effort has been made to increase the diagnostic value of

the present scales to the utmost. The inclusion of a larger number of types than on any other scale of this class permits the identification of disabilities with greater facility. Opportunities for diagnostic work are also inherent in many of the examples though they may not be apparent to a casual inspection. Certain types of errors are more numerous than others and these necessarily determine the form that remedial teaching will take. The variation of the types of errors is so great, however, that it is impossible to give any classification that will serve for all classes beyond a general one that may be fitted to particular needs.

A survey of the most important types of errors was carried out with some of the preliminary scales. For this purpose 150 papers were selected at random for each grade from the second to the eighth, inclusively. The errors were tabulated separately for the four fundamental processes. Classifications of these errors together with the frequency of their appearance have been tabulated in the succeeding pages. Several general results are prominent. The most frequent source of mistakes was in lack of knowledge of the fundamental combinations. No effort has been made to subdivide this class as has been done by others. The second most frequent source of errors involved the mistaking of signs which has an important bearing upon the form of these scales. Signs have been substituted for words in the directions for each example on account of the introduction of a large factor that reduces the validity of such scales when the latter method is employed. The frequency of these errors is greater in the early grades than in the higher ones. The large numbers of pupils who mistook signs at least once reveals the comparative unfamiliarity of such children with the symbols that form an integral part of computation. The use of signs possesses certain limitations but is decidedly superior to the single word method adopted in other mixed fundamental scales. In multiplication and in division much fewer mistakes were made.

Many arbitrary classifications had to be made in recording the frequency of the errors that were committed. This is particularly true in the case of carelessness and guesswork although in many cases it was apparent that the mistakes were due to this cause alone.

Classification of Errors in Addition

	II	III	IV	V	VI	VII	VIII
1. Errors in computation	202	183	74	105	66	78	88
2. Mistaking signs	75	217	107	83	46	40	64
3. Errors in carrying and neglecting to carry	61	3	14	8	5	0	11
4. Guesswork (carelessness in writing answers)	98	14	2	0	14	0	11
5. Failing to complete work	17	6	3	4	0	8	12
6. Failing to reduce answer to lowest terms	0	0	0	8	7	0	4
7. Mistakes in addition of fractions	0	0	0	4	84	22	21
8. Omission of decimal point	0	0	0	0	20	3	5
9. Miscellaneous	27	35	8	0	0	0	0

150 children in each of the above grades.

Classification of Errors in Subtraction

	II	III	IV	V	VI	VII	VIII
1. Errors in combinations	141	206	134	156	117	24	77
2. Mistaking signs	106	126	51	45	15	40	18
3. Errors connected with borrowing	33	91	33	16	23	13	14
4. Inverted digits, e. g., (Subtracted minuend from subtrahend.) 24 −7 / 23	125	163	17	3	0	0	4
5. Incomplete	16	7	14	13	4	3	6
6. Carelessness and guesswork	128	0	15	0	12	0	13
7. Errors in subtraction of fractions	0	0	10	25	25	27	2
8. Subtraction of mixed numbers	0	0	0	0	35	19	9
9. Errors in subtraction of decimals	0	0	1	63	42	9	3
10. Miscellaneous	0	16	0	0	0	1	1

Classification of Errors in Multiplication

	II	III	IV	V	VI	VII	VIII
1. Errors in combinations	132	317	118	175	145	68	136
2. Mistaking signs	66	30	28	37	61	8	21
3. Errors in carrying	13	22	2	15	20	7	15
4. Failure to complete work	12	11	22	27	28	5	9
5. Carelessness and guesswork	70	40	5	0	11	1	13
6. Mistakes in writing of answer	0	6	4	0	5	10	0
7. Errors in multiplying by 0	0	62	29	27	17	11	35
8. Failing to reduce to lowest terms	0	0	5	11	5	4	0
9. Errors in placing of decimal point	0	0	1	63	56	17	24
10. Errors in multiplication of fractions	0	0	1	44	27	25	23
11. Errors in percentage	0	0	0	8	8	9	2
12. Errors in multiplication of mixed numbers	0	0	0	0	35	26	13

Classification of Errors in Division

	II	III	IV	V	VI	VII	VIII
1. Errors in the combinations	77	202	201	140	133	39	132
2. Mistaking signs	14	16	4	0	8	6	9
3. Guesswork and carelessness	128	232	39	0	23	1	10
4. Failure to complete work	4	4	27	43	24	15	22
5. Errors in subtraction	0	32	8	3	6	2	0
6. Errors in dividing 0 by a whole number	22	4	0	0	0	0	0
7. Failure to bring down terms	0	5	24	5	9	9	4
8. Errors in reduction of answer	0	0	16	8	26	18	3
9. Errors in division of fractions	0	0	0	65	92	52	26
10. Errors in decimals and location of decimal point	0	0	2	77	120	117	46
11. Errors in division involving mixed numbers	0	0	0	0	4	14	0

In order to promote the facility with which teachers may classify the errors made by children, a record sheet has been compiled to accompany the scales which contains a list of most of the important errors with spaces left which may be filled in with other types not included in the list. Columns are provided in which the teacher may record the frequency of each type. Those of greatest frequency will naturally be the ones that are most important to overcome. The effects of remedial teaching may be estimated through a repetition of the scales by means of a duplicate form and the errors again tabulated. This use of the scales should recommend them to teachers and should permit large improvements in the ability of the class as a whole and in the efficiency of individual pupils.

In their available form the Diagnostic Computation Scale has two parts, Part I for the second, third and fourth grades, while Part II is primarily intended for the fifth, sixth, seventh and eight grades but may be used in the third and fourth. Accompanying the scales is a manual of directions containing full instructions for their use, the correction answers, and the tables of norms. With each package is supplied, also, a class record sheet for the recording of names and scores and the diagnostic chart. The class record sheet and diagnostic chart are reproduced on the following pages.[29]

[29] The Diagnostic Computation Scales together with all necessary accessories may be obtained from the Catholic Education Press, Brookland Station, Washington, D. C.

The value of these scales will depend upon their conformity to those requirements and principles that have been outlined in the first chapter. An effort has been made to construct them in accordance with these. The ultimate criterion of their worth is the usefulness and accuracy of the results obtained from their use. If, through their employment, a contribution is made to the more efficient teaching of arithmetic, they will have satisfied the major requirements and justified their construction.

Diagnostic Computation Scale

CLASSIFICATION OF ERRORS FOR DIAGNOSTIC PURPOSES

	Types of errors	Frequency	
		1st Test	2nd Test
	A—Integers		
1.	Addition:		
	Mistakes in combinations		
	Neglecting to carry		
	Carrying wrong number		
	Skipping figures in column		
2.	Subtraction:		
	Mistakes in combinations		
	Mistakes in borrowing		
3.	Multiplication:		
	Errors in tables		
	Misplacing right hand figure in partial product		
4.	Division:		
	Errors in tables		
	Error in recording remainder		
5.	Performing wrong operation or mistaking signs.		
	B—Fractions		
1.	Failure to reduce to lowest terms		
2.	Addition:		
	Adding both numerators and denominators		
	Dropping the denominator		
	Adding denominators instead of numerators		
	Not reducing to same denominator		
3.	Subtraction:		
	Omitting the denominator		
	Subtracting denominators		
	Not reducing to same denominator		
	Inability to subtract mixed number from integer		
	Inability to subtract integer from mixed number		
4.	Multiplication:		
	Adding denominators		
	Inability to multiply mixed number by integer		
	Inability to multiply integer by fraction		
	Inverting multiplier		
	Inability to interpret "of"		

Diagnostic Computation Scale—Continued

	Types of errors	Frequency	
		1st Test	2nd Test
5.	Division:		
	Inability to divide fraction by integer		
	Failure to invert divisor		
	Inverting dividend		
	Inverting both dividend and divisor		
6.	Performing wrong operation or mistaking signs		
	C—Decimals		
1.	Addition:		
	Failure to point off sum		
	Misplacing decimal point in sum		
	Errors in addition of decimals when example is written horizontally		
2.	Subtraction:		
	Failure to point off remainder		
	Errors when subtrahend contains more numbers than the minuend		
	Confusing minuend and subtrahend		
3.	Multiplication:		
	Errors in position of decimal point		
4.	Division:		
	Errors in position of decimal point		
5.	Performing wrong operation or mistaking signs		
	D—Miscellaneous		
1.	Failure to complete work		
2.	Guesswork and carelessness		
3.	Errors in percentage		
4.	Errors in operations involving 0		
5.	Mistakes in writing answer or illegible answers		
6.	Performing wrong operation or mistaking signs. (Total of 5A, 6B, and 5C)		
7.	Unclassified		
	E—Types of Errors not Included		
1.			
2.			
3.			
4.			
5.			
	Total		

Diagnostic Computation Scale

CLASS RECORD SHEET

School_____ City_____ State_____

Date_____ Grade_____ Part_____ Form_____ No. of pupils_____

Pupils' Names	Age		Scores		
	Yrs.	Mos.	Ex. C.	Age S.	Ts.
Class Means (Averages)					

BIBLIOGRAPHY

1. Anderson, C. J. The Use of the Woody Scale for Diagnostic Purposes. *Elem. Sch. J.*, xviii, 1918, pp. 770–781.

2. Bird, Grace E. A Test of Some Standard Tests. *J. of Educ. Psy.*, xi, 1920, pp. 275–283.

3. Brown, J. C. A Summary of Some Significant Conclusions Reached by Investigators Relative to Arithmetic. *Elem. Sch. J.*, xxv, 1925, pp. 346–357.

4. Buswell, Guy T., and Judd, Charles H. Summary of Educational Investigations Relating to Arithmetic. *Supp. Educ. Mono.*, No. 27, June, 1925, p. 212. The University of Chicago.

5. Callixta, Sister M. The Educational Objectives of Arithmetic. The Catholic University, *Educational Research Bulletin*, Vol. I, No. 3, 1925, p. 28.

6. Chapman, J. Crosby. A Study of Initial Spurt in the Case of Addition. *J. of Educ. Psy.*, vi, 1915, pp. 419–426.

7. Chapman, J. Crosby, and Nolan, Willis J. Initial Spurt in a Simple Mental Function. *Amer. J. of Psychol.*, xxvii, 1916, pp. 256–260.

8. Clapp, Frank L. The Number Combinations: Their Relative Difficulty and the Frequency of Their Appearance in Textbooks. Univ. of Wisconsin, Bureau of Educ. Research, Bulletin No. 2, July, 1924, p. 120.

9. Clark, John R., and Vincent, E. Leona. A Study of Variability in Arithmetic. *J. of Educ. Psy.*, xvi, 1925, pp. 267–274.

10. Cotton, Carl. The Results of Some Classroom Measurements. *Elem. Sch. J.*, xviii, 1918, pp. 802–806.

11. Counts, George S. Arithmetic Tests and Studies in the Psychology of Arithmetic. *Supp. Educ. Mono.* No. 4, 1917, p. 128. The University of Chicago.

12. Courtis, S. A. Measurement of Growth and Efficiency in Arithmetic. *Elem. Sch. Teacher*, x, 1909, pp. 58–74, 177–199; xi, 1910, pp. 171–185; xii, 1911, pp. 360–370, 528–539.

13. Courtis, S. A. Standard Scores in Arithmetic. *Elem. Sch. Teacher*, xii, 1911, pp. 127–137.

14. Courtis, S. A. The Reliability of Single Measurements with Standard Tests. *Elem. Sch. Teacher*, xiii, 1913, 326–345, 486–504.

15. Courtis, S. A. The Gary Public Schools: Measurement of Classroom Products. New York, General Education Board, 1919.

16. Courtis, S. A., and Thorndike, E. L. Correction Formulae for Addition Tests. *Teachers College Record*, xxi, 1920, pp. 1–24.

17. Finley, G. W. A Comparative Study of Three Diagnostic Arithmetic Tests. Colorado State Teachers College, Bulletin No. 4, Series 20, 1920, p. 40.

18. Gist, Arthur S. Errors in the Fundamentals of Arithmetic. *School and Society*, vi, 1917, pp. 175–177.

19. Guy, J. Freeman. The Pittsburgh Arithmetic Scales. Thesis, Univ. of Pittsburgh, 1923, p. 51.

20. Haggerty, M. E. Arithmetic: A Co-operative Study in Educational Measurements. *Indiana Univ. Studies,* No. 27, *Indiana Univ. Bull.,* Vol. xii, No. 18, pp. 385–507.

21. Haggerty, M. E. Studies in Arithmetic. *Indiana Univ. Studies,* No. 32, 1916, p. 110.

22. Heilman, J. D., and Shultis, F. W. A Study in Addition. *Research Bull.,* No. 1. Colorado State Teachers College, 1916, p. 16.

23. Judd, Charles H. Measuring the Work of the Public Schools. The Cleveland Foundation, Cleveland, Ohio, 1916.

24. Kallom, Arthur W. Analysis of and Testing in Common Fractions. *J. of Educ. Res.,* i, 1920, pp. 177–192.

25. Kallom, Arthur W. The Importance of Diagnosis in Educational Measurement. *J. of Educ. Psy.,* x, 1919, pp. 1–12.

26. Kathleen, Sister M. Some Results of Remedial Instruction Following the Use of Diagnostic Arithmetic Tests. *Catholic Educ. Rev.,* xxiii, 1925, December.

27. Layton, S. H. Some Specific Remedies for Arithmetic Weaknesses. *Univ. of Pennsylvania Bull.,* xx, No. I, 1919, pp. 61–64.

28. Lessenger, W. E. Reading Difficulties in Arithmetical Computation. *J. of Educ. Res.,* xi, 1925, pp. 287–291.

29. McCall, William A. Measuring the Horace Mann Elementary School. *Teachers College Record,* xix, 1918, pp. 472–484.

30. McCarthy, Sister Katharine, O. S. B. A Study of Achievement in the Fundamentals of Arithmetic with Diagnosis of Errors. *Catholic Educ. Rev.,* xxii, 1924, pp. 533–539.

31. Merton, E. L., and Banting, G. O. Remedial Work in Arithmetic. *The Second Yearbook of the Department of Elementary School Principals,* ii, No. 4, 1923, pp. 395–429.

32. Monroe, W. S. An Experimental and Analytical Study of Woody's Arithmetic Scales, Series B. *Sch. and Soc.,* vi, 1917, pp. 412–420.

33. Monroe, W. S. A Series of Diagnostic Tests in Arithmetic. *Elem. Sch. J.,* xix, 1919, pp. 585–607.

34. Monroe, W. S. The Illinois Examination. Univ. of Illinois, *Bureau of Educational Research Bulletin* No. 6, 1921, p. 70.

35. Monroe, W. S. The Ability to Place the Decimal Point in Division. *Elem. Sch. J.,* xviii, 1917, pp. 287–293.

36. Monroe, W. S. The Theory of Educational Measurements. New York, Houghton Mifflin Co., 1923, pp. xxiii+364.

37. Monroe, W. S., DeVoss, J. C., and Kelly, F. J. Educational Tests and Measurements. New York, Houghton Mifflin Co., Revised edition, 1924, pp. xxvii+521.

38. Morton, R. L. An Analysis of Pupils' Errors in Fractions. *J. of Educ. Res.,* ix, 1924, pp. 117–125.

39. Myers, G. C. Persistence of Errors in Arithmetic. *J. of Educ. Res.,* x, 1924, pp. 19–28

40. Myers, G. C. The Prevention and Correction of Errors in Arithmetic. Chicago, The Plymouth Press, 1925, pp. 76.

41. Osburn, W. J. Corrective Arithmetic. New York, Houghton Mifflin Co., 1924, pp. x+182.

42. Osburn, W. J. Errors in Fundamentals of Arithmetic. *J. of Educ. Res.*, v, 1922, pp. 348–349.

43. Osburn, W. J. Improvement in the Fundamentals of Arithmetic. Madison, Wisconsin; State Dept. of Public Instruction, 1923, p. 12.

44. Otis, Arthur S., and Davidson, P. E. The Reliability of Standard Scores in Adding Ability. *Elem. School Teacher,* xiii, 1912, pp. 91–105.

45. Paulu, E. M. Diagnostic Testing and Remedial Teaching. Boston, D. C. Heath and Co., 1924, pp. xviii+372.

46. Phelps, C. L. A Study of Errors in Tests of Adding Ability. *Elem. School Teacher,* xiv, 1913, pp. 29–39.

47. Robertson, M. S. A Diagnostic Fractions Test. *Peabody J. of Education,* i, 1924, pp. 301–310.

48. Ross, Carmon. An Analysis of Pupils' Scores and the Use of Diagnostic Scores in Arithmetic. *Univ. of Pennsylvania Bull.,* xxi, 1921, No. 7, pp. 226–234.

49. Symonds, Percival M. The Accuracy of Certain Standard Tests for School Classification. *J. of Educ. Res.,* ix, 1924, pp. 315–330.

50. Theisen, W. W., and Flemming, Cecile W. The Diagnostic Value of the Woody Arithmetic Scales: A Reply. *J. of Educ. Psy.,* ix, 1918, pp. 475–488, 567–580.

51. Thorndike, Edward L. The Psychology of Arithmetic. New York, The Macmillan Co., 1922, pp. xvi+314.

52. Thorndike, E. L. The Psychology of Drill in Arithmetic: The Amount of Practice. *J. of Educ. Psy.,* xii, 1921, pp. 183–194.

53. Thorndike, E. L. The Constitution of Arithmetical Abilities. *J. of Educ. Psy.,* xii, 1921, pp. 14–24.

54. Trabue, M. R. Measuring Results in Education. New York, The American Book Co., 1924, p. 492.

55. Uhl, W. L. The Use of Standardized Materials in Arithmetic for Diagnosing Pupils' Method of Work. *Elem. School J.,* xviii, 1917, pp. 215–218.

56. Upton, Clifford B. The Influence of Standardized Tests on the Curriculum in Arithmetic. *Teachers College Record,* 1925, pp. 627–641.

57. Wilson, G. M. The Proper Content of a Standard Test. *Elem. School J.,* xix, 1919, pp. 375–381.

58. Wilson, G. M. A Survey of the Social and Business Usage of Arithmetic. *Teachers College Contributions to Education,* No. 100, 1919, pp. vi+64.

59. Wilson, G. M., and Hoke, Kremer J. How to Measure. New York, The Macmillan Co., 1921, pp. vii+285.

60. Wilson, W. K. The Woody-McCall Arithmetic Test as a Test in Reading. *Educational Research Bulletin,* iii, 1924, p. 329.

61. Wood, E. R. Investigations in Arithmetic. *Kentucky High School Quarterly,* v, 1919, No. 4, pp. 1–49; No. 5, pp. 1–58.

62. Woody, Clifford. Measurements of Some Achievements in Arithmetic. *School and Society,* iv, 1916, pp. 299–303.

63. Woody, Clifford. Measurements of Some Achievements in Arithmetic. *Teachers College Contributions to Education,* No. 80, 1920, revised edition, p. 78.

64. Woody, Clifford. Results from Successive Repetitions of Certain Arithmetic Tests. *Bulletin of the Bureau of Education,* University of Indiana, Vol. I, No. 3, 1925, pp. 61–79.

NCP
NATIONAL CAPITAL PRESS, INC., WASHINGTON, D. C.